In silico expression analysis to identify potentially functional plant *cis*-regulatory elements

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina

zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Julio Cesar Bolivar Lopez
aus Bogotá

1. Referent: apl. Professor. Dr. Reinhard Hehl
2. Referent: Professor Dr. Dieter Jahn
eingereicht am: 10.06.2013
mündliche Prüfung (Disputation) am: 14.10.2013

Druckjahr 2013

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Hehl, R., J. C. Bolivar, J. Koschmann, Y. Brill, L. Bülow (2013) Databases and web-tools for gene expression analysis in *Arabidopsis thaliana*. In: Advances in Genome Science: Probing intracellular regulation (Volume 2). Neri C, ed. Bentham eBooks.

Tagungsbeiträge

Bolivar, J.C., Pajonk, S., Krämer, U., Roccaro, M., Somssich, I. & Bülow, L.. Prediction of novel stress-related *cis*-regulatory elements. GABI Status Seminar. Postdam, Germany (2010).

Bolivar, J.C., Brill, Y., Hehl, R. & Bülow, L. *In silico* expression analysis to identify potentially functional plant *cis*-regulatory elements. Plant & Animal Genome XIX. San Diego, United States. (2011).

Bolivar, J.C., Pajonk, S., Krämer, U., Roccaro, M., Somssich, I. & Bülow, L. Prediction and validation of novel endogenous and synthetic stress-related *cis*-regulatory elements. 11. Status Seminar. Potsdam, Germany. (2011).

Bolivar, J.C., Brill, Y., Hehl, R. & Bülow, L. Novel endogenous and synthetic stress-responsive *cis*-regulatory elements. Plant 2030 Status Seminar. Potsdam, Germany. (2012).

So much, after all, remains in our thoughts and hearts

as unrealized suggestion.

Andrei Tarkovsky

(1932-1986)

Table of Contents

List of Abbreviations

ABA – abscisic acid

ABRE – abscisic acid responsive element

AGI – *Arabidopsis* genome initiative

ANR – any number of repetitions

API – application programming interface

CBF – c-repeat binding factor

ChiP – chromatin immunoprecipitation

CRE – *cis* regulatory element

CREF – *cis*-regulatory element finder

EF-Tu – elongation factor thermo unstable

EF1A – elongation factor 1 alpha

EM – expectation maximization

DRE – dehydration responsive element

DREB1 – dehydration-responsive element binding protein 1

Flg – flagellin

JAR – java archive

LUC – luciferase

MAPK – mitogen-activated protein kinase

MEME – multiple expectation maximization for motif elicitation

NPP1 – necrosis-inducing *Phytophthora* protein 1

OOPS – one occurrence per sequence

P. syringae – *Pseudomonas syringae*

PAMP – pathogen-associated molecular pattern

Pb – lead

PSSM – position-specific scoring matrix

PWM – positional weight matrix

SOTA – self-organizing tree algorithm

SQL – structured query language

synCRE – synthetic *cis*-regulatory elements

TAIR – *Arabidopsis* information resource

TF – transcription factor

TFBS – transcription factor binding site

TMV – tobacco mosaic virus

TSS – transcription start site

UPGMA – unweighted pair group method with arithmetic mean

ZDRE – zinc deficiency responsive element

Zn – zinc

ZOOPS – zero or one repetition per sequence

1 Introduction

1.1 Gene expression and *cis*-regulatory elements

Eukaryotic gene expression is largely controlled by the binding of transcription factors to *cis*-regulatory elements (CREs) in promoter regions. Such CREs, typically 8 to 10 nucleotides long and the proteins binding to them, differ among eukaryotic genes. In the model plant *Arabidopsis thaliana* approximately 1346 to 2290 genes have been reported to express putative transcription factors (Davuluri et al. 2003; Guo et al. 2008). The binding sites can be located farther away from the translation start site (TSS) and still have an effect on gene expression through DNA looping (Schleif 1992). That process facilitates interactions between regulatory proteins very near to the TSS (the so called general transcription factors) and regulatory proteins binding farther away to the TSS (Schleif 1992). These interactions often occur with the help of mediator proteins (Lee and Young 2000). A simplified view of a eukaryotic gene is presented in **Figure 1.1** where the interactions between CREs, gene regulatory proteins, general transcription factors and mediator proteins are illustrated.

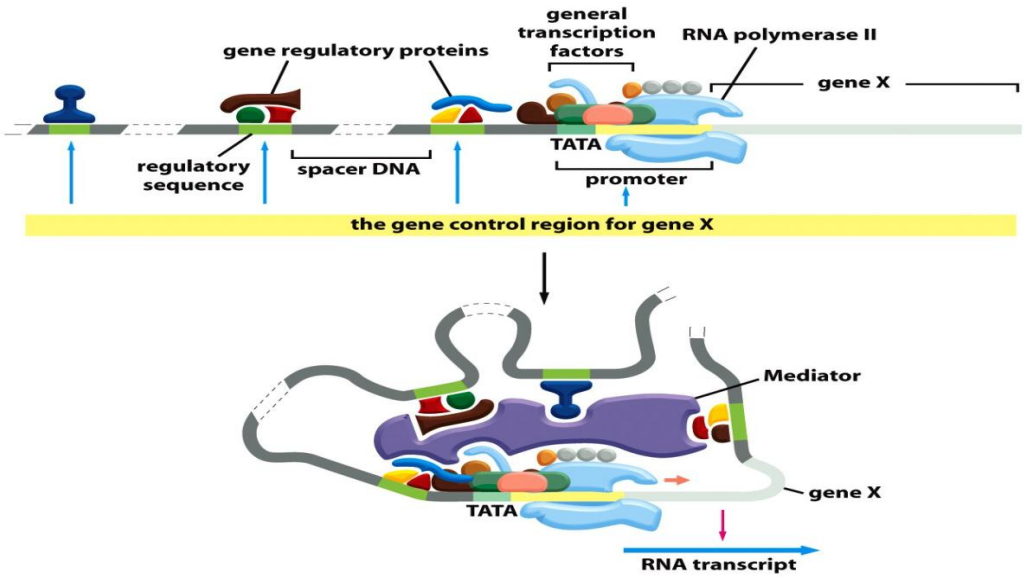


Figure 1.1: Simplified view of a eukaryotic gene showing its promoter region with *cis*-regulatory elements and TATA-Box binding sites. CREs located upstream of TSS serve as binding sites for transcription factors that act as regulatory proteins, controlling the expression of the gene. Modified from (Alberts 2008).

Proteins regulating gene expression can specifically recognize bases from the CREs which are often exposed in the major groove of the DNA strand (Pabo and Sauer 1992). The recognition occurs due to a contact and interaction between the surface of the protein and the exposed surface of the DNA in that region (Pabo and Sauer 1992). These interactions constitute hydrogen bonds, ionic bonds and hydrophobic interactions, which together confer a very strong stability of the protein-DNA interaction (Pabo and Sauer 1992). The proteins display several structural motifs that are used for the interactions with the DNA, some of the most common structural motifs include the Helix-Turn-Helix, Homeodomain, Zinc-finger and ß-sheets among others (Pabo and Sauer 1992).

The process of eukaryotic expression regulation requires a large number of proteins including regulatory proteins, RNA-polymerase II and general transcription factors (Roeder 1996). The regulatory proteins are the ones that activate or inhibit transcription of a given gene (Roeder 1996). In order to activate gene transcription, activator proteins facilitate the assembly of the RNA-polymerase II and general transcription factors upstream to the TSS, i.e. within the promoter region (Roeder 1996). This assembly process can occur as a result of different mechanisms, e.g. the activator proteins can promote the rapid binding of the general transcription factors to the promoters or they can attract the mediator protein in order to enable binding of RNA-polymerase II and general transcription factors binding and in that way initiate transcription (Green 2005). CREs are short DNA-sequences generally located within the promoter regions of genes that can specifically been recognized by transcription factors. Several functional CREs have been reported for plants, some well-known examples will be presented in the next chapter. Often activator proteins work synergistically with other activator proteins, this topic will be covered in **Chapter 1.2**.

1.1.1 Biotic and abiotic stress responsive *cis*-regulatory elements in plants

Since plants are sessile organisms, they have developed sophisticated mechanisms to cope with environmental stresses. It is essential for the plant to adapt to a whole array of biotic and abiotic stresses by altering metabolism and growth as well as by expression of specific resistance and tolerance proteins. This expression is mainly

achieved by transcriptional gene activation. Several sequences have been reported as CREs for plants (Davuluri et al. 2003; Guo et al. 2008; Higo et al. 1999; Bülow et al. 2010). The present work aims at predicting novel CREs, therefore, to gain an insight about the relevance and functionality of CREs, some well-characterized CREs will be presented in this chapter.

The abscisic acid (ABA) signaling pathway is well known for plant responses towards environmental stresses. A large number of genes involved in the ABA signaling pathway have been identified and it has also been shown that the regulation of such genes is controlled by conserved promoter sequences which serve as binding sites of regulatory proteins (Yoshida et al. 2010). These CREs are known as Abscic Acid Responsive Elements (ABREs) and they have been demonstrated to be involved in ABA responses under abiotic stresses (Yoshida et al. 2010; Kim et al. 2011). Functional ABREs have been reported for rice and for *Arabidopsis* (Choi et al. 2000), highlighting the importance of such CREs for the control of gene expression in stress responses.

Another very important CRE involved in plant abiotic stress-responses is the Dehydration Responsive Element (DRE). This CRE has been demonstrated to be involved in fast responses towards dehydration, high salinity and cold stresses (Nakashima et al. 2009). The importance of the element was demonstrated by showing that no other element is required for the dehydration response (Yamaguchi-Shinozaki and Shinozaki 1994). Furthermore, the element has been shown to be important for the binding of different transcription factors in *Arabidopsis*, including the Dehydration-responsive element binding protein 1 (DREB1), the C-repeat binding factor (CBF) and DREB2, which control the plant responses towards abiotic stresses (Nakashima et al. 2009).

One of the best-characterized plant CRE responsive to biotic stresses is the W-box. It is known that the sequence serves as binding site of WRKY transcription factors which have been extensively demonstrated to be involved in plant immune responses (Rushton et al. 2010; Pandey and Somssich 2009) and defense signaling (Eulgem and Somssich 2007). The W-boxes are highly related to each other and are characterized by the core sequence TGAC, which is required for the specific binding of WRKY

transcription factors (Rushton et al. 2010). Although single CREs are very important for gene expression regulation, this process often occurs cooperatively, where several CREs serve as binding sites of proteins that interact with each other to produce a response. It has been shown that sequences adjacent to W-boxes are also required for gene transcription (Rushton et al. 2010). The topic of combinatorial regulation will be covered in the next chapter.

1.2 Transcriptional regulation through combinatorial control

Eukaryotic gene activator proteins work synergistically to initiate transcription of a given gene. Combinations of proteins that bind to specific CREs in gene promoters can cause high increases in transcription rates (Carey 1998). This combinatorial effect is much higher than the expected effect from the sum of the single elements, a process called transcriptional synergy (Carey 1998). Thus, the specific sequences of the CREs guide the assembly of the proteins that will activate transcription (Levine and Tjian 2003). These activator proteins bind first to chromatin and are then organized into a complex called the enhanceosome (see **Figure 1.2**) (Carey 1998). This enhanceosome is formed by a large number of sequence-specific activator proteins that interact with DNA-binding proteins (Carey 1998). This complex interacts with coactivators through protein-protein interactions resulting in the recruitment of the RNA-Pol II and other factors to form the preinitiation complex, which ultimately leads to synergistic effects on transcription (Carey 1998). Examples of combinatorial control have been reported for different species (Kel et al. 1995; Yuh et al. 1998; Wang et al. 1999).

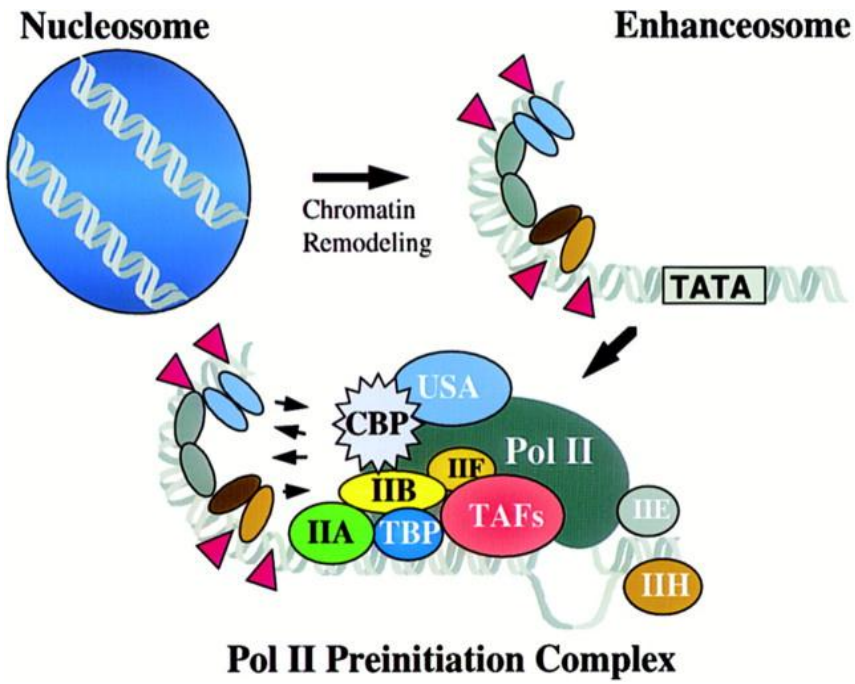


Figure 1.2: Eukaryotic enhanceosome and preinitiation complex. Sequence-specific activators are shown as ovals and DNA-bending proteins as triangles (Carey 1998).

In plants it has been shown that the expression of genes induced by ABA is mediated through combinatorial control in barley (Singh 1998). A combination of an ABRE and a coupling element (CE) constitute a combinatorial ABRE (ABREC1) (Singh 1998). The importance of the spacer separating two CREs is shown in **Figure 1.3**. A combination (ABRC1) is given with ABRE A3 and CE1 that display a spacer length of 20 nucleotides (Singh 1998). On the other hand, another combination (ABRC3) with CE3 and ABRE A2 shows no spacer (Singh 1998). This combination serves to attract the protein VP1 which then enhances transcription (Singh 1998).

The spacer between combinatorial CREs has been reported to be very important in other studies, where it was shown that a protein can interact with different kinds of proteins depending on the length of the DNA spacer (Reményi et al. 2004). For plants, databases such as PlantPAN (Chang et al. 2008) and web services from AthaMap (Steffens et al. 2005) emphasize the importance of the spacer length for combinatorial CREs. In *Saccharomyces cerevisiae,* analyzing predicted combinatorial elements, further spatial constraints, such as relative orientation to each other and to the direction of transcription, were also demonstrated to have an effect on the combinatorial elements functionality (Yu et al. 2006). All these spatial constraints

dictate the precise arrangement of regulatory proteins in gene promoters which will ultimately control the expression of genes under different conditions (Singh 1998).

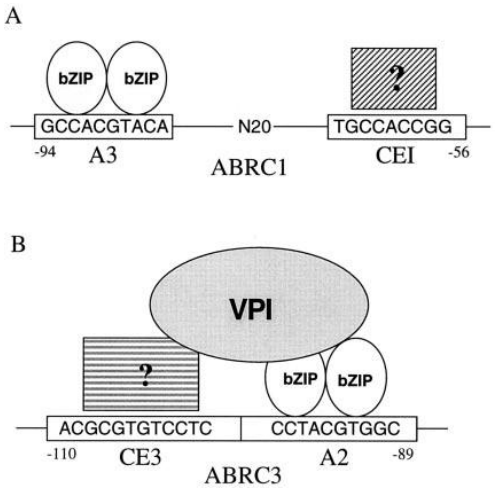


Figure 1.3: Two combinations of CREs (ABRC1 and ABRC3) responsive to ABA. ABRC1 and ABRC1 are formed by the ABREs A3 and A2 respectively, which are known binding sites of bZIP proteins. Proteins binding to CE1 and CE3 are not identified (Singh 1998).

Until this part, only examples have been given were CREs serve as binding sites of transcription factors that respond to a single stress. However, in nature stress responses overlap giving rise to crosstalks, a topic that will be covered in the next chapter.

1.3 Crosstalks between abiotic and biotic stresses

Like in other organisms, stress signaling in plants is mainly transduced via MAP kinase (MAPK) cascades. These are involved in responses to various biotic and abiotic stresses, but also associated with hormone signaling and cell division and developmental processes. The fact that the Arabidopsis genome harbours only 20 MAPKs, 10 MAPK kinases and 60 MAPK kinase kinases (Group M. 2002; Rodriguez et al. 2010) strongly implies a necessity for the plant to converge signaling pathways at this bottleneck. The existence of pathway crosstalks in plants has been previously reported (see **Figure 1.4**). Several transcription factors and kinases have been identified as possible candidates having a role in different signaling pathways (Fujita et al. 2006). For example, it has been shown that crosstalks occur between salicylic acid, jasmonic acid, ethylene and other phytohormones signaling pathways (Bostock 2005). In

Arabidopsis, jasmonic acid and salicylic acid lead to synergistic effects in responses to bacterial pathogens (Bostock 2005). Further evidence has been presented to demonstrate that signaling networks related to abiotic stress tolerance and disease-resistance significantly overlap (Mauch-Mani and Mauch 2005).

Also in *Arabidopsis,* it has been shown that the biotic and abiotic stress responses overlap (Narusaka et al. 2004). It was demonstrated that cytochrome p450 genes were expressed under biotic and abiotic stresses (Narusaka et al. 2004). Such genes contain MYB, TGA-Box and W-box promoter binding sites (Narusaka et al. 2004), indicating that similar transcription factors can bind to these gene promoters and act as signaling convergence points.

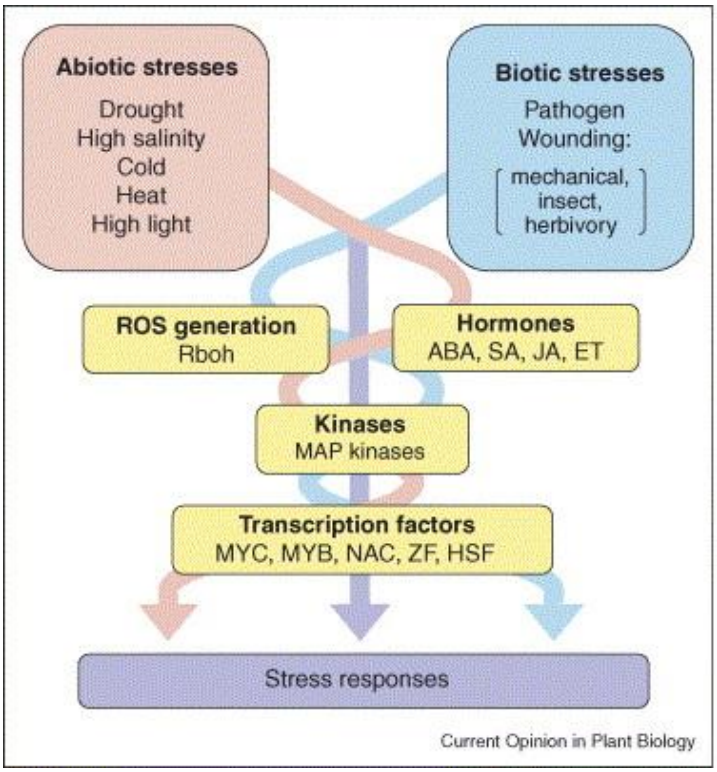


Figure 1.4: Abiotic and biotic signaling pathway. Possible convergence points where pathways crosstalk are also shown (Fujita et al. 2006).

As already mentioned, a common point where several plant signaling pathways converge is at the well-known MAPK cascades (Chinnusamy et al. 2004). MAPKs are involved in a wide range of signaling pathways, such as biotic, abiotic, developmental and hormonal stress signaling, which indicates that MAPK cascades are a convergence point in stress signaling (Chinnusamy et al. 2004). For example MAPK cascades have

been shown to be activated in rice under ABA, biotic and abiotic stresses, including wounding, drought, salt and cold stresses (Chinnusamy et al. 2004).

Further downstream of the MAPK cascades, transcription factors are activated in order to regulate gene expression. 40 genes encoding transcription factors were identified, that were inducible by drought, cold or high salinity stresses, suggesting the existence of common regulatory mechanisms for these stresses TF family members (Seki et al. 2002). Among the common transcription identified were factors DREB, ERF, WRKY, MYB, bHLH, bZIP and NAC (Seki et al. 2002). Osmotic stresses have also been reported to be involved in pathway crosstalks (Chinnusamy et al. 2004). Further common points of convergence were shown with ABA stresses which lead to the activation of stress responsive genes through the binding of transcription factors to DRE, MYB and ABRE CREs (Chinnusamy et al. 2004). The transcription factor DREB2A displays a dual function under drought and high temperature stresses (Qin et al. 2011). Under drought, drought and heat or heat stress different genes are regulated by the binding of the DREB2A protein to DRE CREs within gene promoters, which enhances the plant's tolerance towards these stresses (Qin et al. 2011).

Thus, overlapping gene sets are activated or inhibited under biotic and abiotic stresses (Fujita et al. 2006). The large number of plant signaling networks crosstalk among each other, giving the plant common regulatory mechanisms to respond to different stress types through specific gene regulation (Fujita et al. 2006). DNA microarrays have been very important for the elucidation of such mechanisms, giving the opportunity to measure the expression of thousands of genes under different conditions and thereby allowing to determine crosstalk between different pathways (Ma et al. 2006). Details about how these microarrays are used to assess gene expression will be given in the next chapter.

1.4 DNA microarrays to measure gene expression levels

DNA microarrays, developed in the 1990s, are used to monitor the expression of many genes at once (Lockhart and Winzeler 2000). The first DNA microarray for a eukaryotic genome was developed in 1997 for the yeast *Sacharomyces cerevisae* (Lashkari et al.

1997). These arrays facilitate the identification of gene expression patterns by showing which genes are being expressed and repressed in a certain time point and condition (Lockhart and Winzeler 2000).

The microarrays are small glass slides with specific and known DNA fragments that are automatically attached to its surface and serve as reporters (or probes) (Lockhart and Winzeler 2000). In order to monitor gene expression, mRNAs from two samples (experimental and reference one) are isolated and converted to cDNA by a reverse transcriptase. The DNA in both samples is fluorescently labeled, one with red fluorochrome and one with green fluorochrome (Lockhart and Winzeler 2000). The first microarrays developed were the so-called cDNA arrays where whole cDNAs are spotted onto the glass slides. The samples are allowed to hybridize with the DNA in the microarray and after a period of incubation, the microarray is washed thereby removing weakly bound cDNA (Lockhart and Winzeler 2000). The process occurs, since cDNA strands that are completely complementary with the DNA in the array will form a stable bond, whereas this bond with partially complementary strands will be weak and therefore be washed away. The fluorescence in the array is scanned to reveal the expression of the genes in the experimental sample compared to the expression in the reference sample (Lockhart and Winzeler 2000). Thus, red spots in the array mean that the expression of a given gene is higher in the experimental sample, whereas green spots mean the expression is lower than in the reference condition sample (Lockhart and Winzeler 2000).

Another microarray technology is the oligonucleotide arrays from the company Affymetrix®. These arrays are designed using sequence information alone and are constructed by *in situ* light-directed oligonucleotide synthesis using two procedures: photolithography and solid-phase DNA synthesis (Lipshutz et al. 1999). For this purpose protected photochemically modified linkers are attached to a glass surface (Lipshutz et al. 1999). Selected sites are deprotected and activated by directing light through a photolithographic mask (Lipshutz et al. 1999). Protected nucleotides will couple to these activated sites and the process will be repeated which permits the construction of DNA probes (Lipshutz et al. 1999). This type of array allows a very high density of probes enabling parallel analysis of thousands of genes. A cDNA sample

fluorescently tagged will hybridize to complementary probes on the array (Lipshutz et al. 1999). The fluorescence is then measured to assess levels of gene expression (Lipshutz et al. 1999). When using oligonucleotide arrays, the reference cDNA sample is always hybridized on a separate array.

Once the fluorescence in a microarray has been measured, the information can be used to determine gene sets sharing similar expression profiles by a method called *cluster analysis* (Eisen et al. 1998). Such clustered gene sets are expected to be co-regulated and expressed under similar conditions giving hints about shared functionality (Eisen et al. 1998). Different types of commercially developed microarrays are available (Bammler et al. 2005) and one of the most popular ones is the Affymetrix GeneChip (Irizarry et al. 2003).

DNA microarrays have been extensively used in plants to measure gene expression and to identify co-regulated genes under several conditions (Aharoni and Vorst 2002). A study reported the expression of profiles of several *Arabidopsis* genes under drought, cold and high salinity stresses (Seki et al. 2002). With the help of DNA microarrays, 22 genes were identified as being responsive to all analyzed stresses (drought, cold and high salinity) (Seki et al. 2002). Using DNA microarrays, *Arabidopsis* expression profiles of genes being expressed during seed development were used to identify approximately 650 genes being differentially expressed (Girke et al. 2000). *Arabidopsis* genes regulated under diurnal and circadian cycles were analyzed with microarrays to find genes regulated exclusively under the circadian clock (Schaffer et al. 2001). DNA microarrays have also been used for plant defense analysis in *Arabidopsis*, where the finding of differentially expressed genes after treatment with a fungal pathogen led to the conclusion that salycilic and jasmonate signaling pathways act in an antagonistic manner (Schenk et al. 2000).

DNA microarrays can be used to identify genes being up- or down-regulated upon a given condition. A very common technique to find possible common regulators of such genes is to identify conserved sequences in their promoters with the help of motif prediction programs. In the present study the well-known MEME algorithm was used for motif prediction, it will be described in the next chapter.

1.5 Prediction of *cis*-regulatory element motifs with the MEME algorithm

Motifs can be described as short similar sequence patterns occurring several times within a set of sequences and which may serve as binding sites for transcription factors (Das and Dai 2007). The problem of finding such a motif within input sequences is solved by searching for statistically overrepresented motifs among input sequences (Das and Dai 2007). One very common computer algorithm for motif discovery is MEME (Bailey and Elkan 1994) whose name stands for Multiple Expectation Maximization for Motif Elicitation.

The MEME algorithm identifies non-overlapping motifs without insertions or deletions present in a set of input sequences (Bailey and Elkan 1995b). The algorithm does not make any assumptions about the position or the number of motifs present in the input data (Bailey and Elkan 1995b). MEME includes novel methods for motif discovery that will be explained next. First, shorter sequences present in the input sequences are used as starting points and with the help of an expectation maximization (EM) algorithm, local optimal motifs are found (Bailey and Elkan 1995b). The EM algorithm estimates the model parameters thus maximizing the likelihood of the observed data (Bailey and Elkan 1994). Another novel method employed in MEME to manage noise in the input data is the possibility to modify the EM algorithm so that motifs can occur zero, one or multiple times (Bailey and Elkan 1995b). For that purpose MEME can find different types of motif occurrences, the simplest one is “one occurrence per sequence” (OOPS), which is also the one that takes least computational time (Bailey and Elkan 1995a). Another one is the generalized version of OOPS, which is “zero or one repetition per sequence” (ZOOPS) and in turn takes more computational time than OOPS (Bailey and Elkan 1995a). The last possible type of motif occurrences is “any number of repetitions” (ANR) which allows zero, one or several non-overlapping motif repetitions in the input sequences, ANR is the option that takes most computational time (Bailey and Elkan 1995a). Finally, one very important feature of MEME is its ability to mask motifs that have already been found, thereby allowing the identification of a large number of distinct motifs (Bailey and Elkan 1995b). That is accomplished by performing a so called greedy search, which includes information of the motifs that

have been discovered into the model thereby avoiding rediscovery of a previously found motif (Bailey and Elkan 1995a).

Being a very important field in bioinformatics, there are several other available motif finding programs, making the decision of using a program over another rather difficult (Sandve and Drabløs 2006). Therefore, benchmarking tests have been developed in order to facilitate comparison among different motif finding programs, a topic that will be covered in the next chapter.

1.5.1 Benchmarking tests of available motif finding programs

Several algorithms have been developed to date for the task of finding conserved motifs among a certain number of input sequences (Das and Dai 2007). In order to compare different motif finding algorithms, two important benchmarking tests have been performed (Tompa et al. 2005; Hu et al. 2005). The aim of the first test was to serve as a guidance to evaluate the accuracy of motif finding algorithms as well as to generate a benchmark data set that will facilitate the assessing of further algorithms (Tompa et al. 2005). For the test, test data were created which contained known binding sites from the TRANSFAC database (Tompa et al. 2005). Several motif finding programs (including MEME) were used to find one over-represented motif in the input data (Tompa et al. 2005). The study reported the program Weeder (Pavesi et al. 2004) outperforming the other programs (Tompa et al. 2005). However, as noted by the authors, the study is a first attempt with space for improvement (Tompa et al. 2005). Several criteria were proposed to improve the benchmarking test, with the most important one being the possibility to predict more than one motif, i.e. not only take the 'best' hit predicted by a program (Tompa et al. 2005). In reality, the top predictions of each motif finding program are actually pursued, rather than relying on the first prediction (Tompa et al. 2005). Another very important test was developed which took into account the fact that the 'best' hit reported by the motif finding algorithms is not always the most accurate prediction (Hu et al. 2005).

The performance of the widely used motif finding programs AlignAce(Hughes et al. 2000), MEME (Bailey et al. 2009), BioProspector (Liu et al. 2001), MDScan (Liu et al.

2002) and MotifSampler (Thijs et al. 2002) was evaluated with another benchmark test (Hu et al. 2005). The test defined performance indexes that were applied to evaluate accuracy, scalability and reliability of the motif predictions (Hu et al. 2005). Features, such as the motif width, input sequences number and motif information content were assessed, in order to determine their influence on motif prediction (Hu et al. 2005). An index called the motif level success rate (mSr) was developed to evaluate the accuracy of the motif finding programs (Hu et al. 2005). In addition, the mSr of MEME was the highest among the tested algorithms (Hu et al. 2005). Interestingly, the test showed that the top-scoring motif predicted by the programs was not the best prediction (Hu et al. 2005), this significant result highlights the importance of pursuing more top predictions from a motif finding program. The algorithms scalability, i.e. coping with varying numbers of input sequences, different motif widths and sequence lengths, was measured in order to determine how it affects the performance of the algorithm (Hu et al. 2005). Among the factors affecting motif prediction accuracy, the length of the input sequences was shown to be the most important one (Hu et al. 2005). Notably, it was demonstrated that when the length of the input sequences is increased the predictions drop significantly (see **Figure 1.5**). In addition, the figure shows that MEME was the program that best performed in the prediction of motifs with respect to different input sequence lengths, a result proposed to be the effect of the high sensitivity of MEME (Hu et al. 2005). All motif finding programs allow the definition of the motifs width to be predicted. With MEME it is further possible to define a range of desired motif widths, a feature that was shown to be very important for motif prediction accuracy (Hu et al. 2005). Another very important characteristic observed in the study was the fact that adding more than 40 motif containing sequences does not improve the accuracy of the predictions (Hu et al. 2005), which may serve as a guideline when the programs are used.

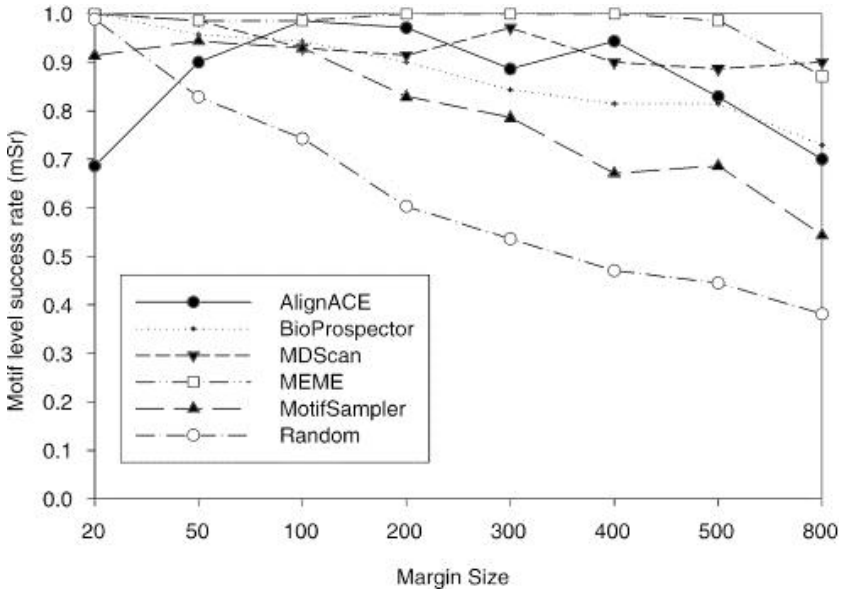


Figure 1.5: Effect of sequence length on motif prediction accuracy (Hu et al. 2005).

Motif finding programs generate a very large number of predictions. The time and resources needed to experimentally validate such a large number of predictions makes the task very difficult to accomplish. For this reason a new motif validation tool was developed in this study. The tool calculates the probability that a given sequence or motif is a functional CRE. In this way the tool allows to reduce the number of predictions performed with motif finding programs by selecting the motifs having a high probability of being functional. Such predictions can be assessed for novelty by using biological databases. In the next chapters the databases used throughout this study will be described.

1.6 Biological databases

In Bioinformatics, it is indispensable to process and analyze large amounts of third-party data. For bioinformaticians working with plants, there are numerous public biological databases on plant genomics, transcriptomics, transcription factors and *cis*-elements (Hehl and Bülow 2008). Four databases which were used extensively in this study were PathoPlant, AthaMap, Place and AGRIS. Information about *Arabidopsis* gene expression can be retrieved for example from the PathoPlant database. PathoPlant was initially developed as a relational database containing molecules and reactions related to Plant-Pathogen interactions as well as molecules involved in signal

transduction pathways related to plant pathogenesis (Bülow et al. 2004). Microarray expression data is also stored in PathoPlant (Bülow et al. 2007). With the help of web tools implemented in the database, gene expression data is used to identify genes up- or down-regulated upon certain stresses (Bülow et al. 2007). Genes identified as being responsive to a certain stress can be further analyzed with the AthaMap database, which highlights the integration of both databases (Bülow et al. 2007). PathoPlant is publicly available at http://www.pathoplant.de.

The AthaMap database was developed as a resource for binding sites of transcription factors in the genome of *Arabidopsis thaliana* (Steffens et al. 2004). These binding sites were predicted by screening alignment matrices and single sequences corresponding to transcription factor binding sites in the *Arabidopsis* genome (Steffens et al. 2004, Bülow et al. 2006). The predicted binding sites can be accessed via absolute positions in a genome and with *Arabidopsis* Genome Initiative identification numbers (AGIs) (Steffens et al. 2004). The database features several tools for different purposes, such as combinatorial element prediction (Steffens et al. 2005), gene analysis for the discovery of common CREs, (Galuschka et al. 2007), identification of post-transcriptionally regulated genes (Bülow et al. 2009), the identification of all binding sites of a specific transcription factor within a specific range (Bülow et al. 2010) and the identification of MicroRNA targets (Bülow et al. 2012). AthaMap is freely accessible at http://www.athamap.de.

The database PLACE provides an extensive collection of plant CREs that have been imported from published literature (Higo et al. 1999). The original sequence of these CREs as well as reported variations in other genes and species are provided (Higo et al. 1999). From the web interface it is possible to perform keyword and signal searches (Higo et al. 1999). Keyword searches can be carried out for motif names, stress types, tissues, sequence and plant species, among others (Higo et al. 1999). The search using signal scan identifies identical or similar CREs to the ones provided by the users (Higo et al. 1999). The database is available at http://www.dna.affrc.go.jp/PLACE/.

The Arabidopsis Gene Regulatory Information Server (AGRIS) stores information about transcription factors and CREs (Yilmaz et al. 2011). AGRIS is primarily focused on

Arabidopsis and experimental data related to transcription factors and gene regulatory networks (Yilmaz et al. 2011). Two databases comprise AGRIS, AtTFDB and AtcisDB (Yilmaz et al. 2011). AtTFDB stores information about transcription factors and their families on the base of the presence of conserved domains (Yilmaz et al. 2011). AtcisDB contains information about genes up-stream regions containing CREs, thereby mapping the CREs to their corresponding promoter locations and also making a distinction between predicted and experimentally validated sites (Yilmaz et al. 2011). All information of AGRIS is publicly available at http://arabidopsis.med.ohio-state.edu/.

1.7 Software

A description of the software used in this study will be given in the next chapters. First the installed version of the motif prediction program MEME will be explained in **Chapter 1.7.1**. Then, details about the web server STAMP and its function for comparative DNA analysis will be given in **Chapter 1.7.2**. All software developed in this study was written in the Java programing language which is shortly described in **Chapter 1.7.3**. Finally the general architecture of MicrosoftSQL, the relational database management system of PathoPalnt and AthaMap is explained in **Chapter 1.7.4**.

1.7.1 MEME: Multiple EM for Motif Elicitation

MEME is used for the finding of statistically significant overrepresented motifs within a set of input sequences (Bailey et al. 2006). Details about the algorithm underlying MEME were explained in **Chapter 1.5**, and therefore this chapter is about the locally installed version of MEME. Academic users can freely download MEME at the website http://meme.sdsc.edu/meme/meme-download.html where details about the installation are available. To use MEME the users have to submit a set of sequences in the FASTA format (Bailey et al. 2006). Such sequences should be preferably short (ideally less than 1000bp long) and sequences believed to not contain the expected motifs should also be filtered out (Bailey et al. 2006). Users are also advised not to include more than 40 input sequences, since a higher number does not improve discovery of motifs (Bailey et al. 2006). Also by default, MEME automatically chooses

motif width (between 6 and 50) and the number of minimum and maximum motif occurrences, although all these parameters can also be changed by the users (Bailey et al. 2006). The output produced by MEME contains all the information about the predicted motifs, including their matrices (derived from the multiples alignment) and their relative positions within the input sequences in the form of 'Block diagrams' (Bailey et al. 2006).

1.7.2 STAMP web server

Motif-finding programs like MEME can predict a high number of statistically significant motifs. Researchers often face the problem of determining the novelty and similarities between such computationally predicted motifs and known motifs from biological databases. A very useful resource for determining motif similarity and novelty is the STAMP web server (Mahony and Benos 2007). **Figure 1.6** summarizes the STAMP web server steps for motif analysis. STAMP accepts user-submitted motifs and queries them against certain user-selected biological database (Mahony and Benos 2007). In order to compare the similarities between two motifs, STAMP performs an alignment using the Needleman-Wunsch, the Smith-Waterman or a special ungapped type of the Smith-Waterman algorithm, which is the default option (Mahony and Benos 2007). The columns in the alignment are compared with scores determined by some of the supported distance metrics, the default being Pearson Correlation Coefficient (Mahony and Benos 2007). In addition there are several options for the gap-opening penalty which differ depending on the metric used to compare columns (Mahony and Benos 2007). STAMP provides a familial profile of the motifs entered as input (see **Figure 1.6**) which is performed with a multiple motif alignment (Mahony and Benos 2007). Two alignment strategies can also be selected for this purpose: progressive profile alignment, where motifs are added in order of decreasing similarity, and iterative refinement (the default option), where, after having identified and aligned the two most similar motifs, motifs are added according to their similarity to the current alignment (Mahony and Benos 2007).

STAMP also constructs similarity trees with the input motifs (see **Figure 1.6**) which are built depending on the selected option: unweighted pair group method using

arithmetic averages (UPGMA) or self-organizing tree algorithm (SOTA) (Mahony and Benos 2007). UPGMA (default option in STAMP) first assigns a node to each input motif, then, it clusters nodes together by identifying the node with the maximum average pairwise similarity (Mahony and Benos 2007). SOTA on the other hand starts from a root node, which corresponds to the alignment of all input motifs, this root node produces two identical leaf nodes and from them SOTA assigns new nodes until each leaf is comprised of a single node (Mahony and Benos 2007). All the input motifs can be queried against user provided motifs or against a selected database, where the top 1, 5 or 10 best matches are presented as result (Mahony and Benos 2007). In the case of plants three databases are available: AthaMap, Agris and PLACE. In addition several input data formats are supported by STAMP, including the output produced by the program MEME, among others (Mahony and Benos 2007). Finally, STAMP trims by default the motif edges with low information content and the users can select any combination of the previously mentioned search parameters (Mahony and Benos 2007). The STAMP web server is available at http://www.benoslab.pitt.edu/stamp/.

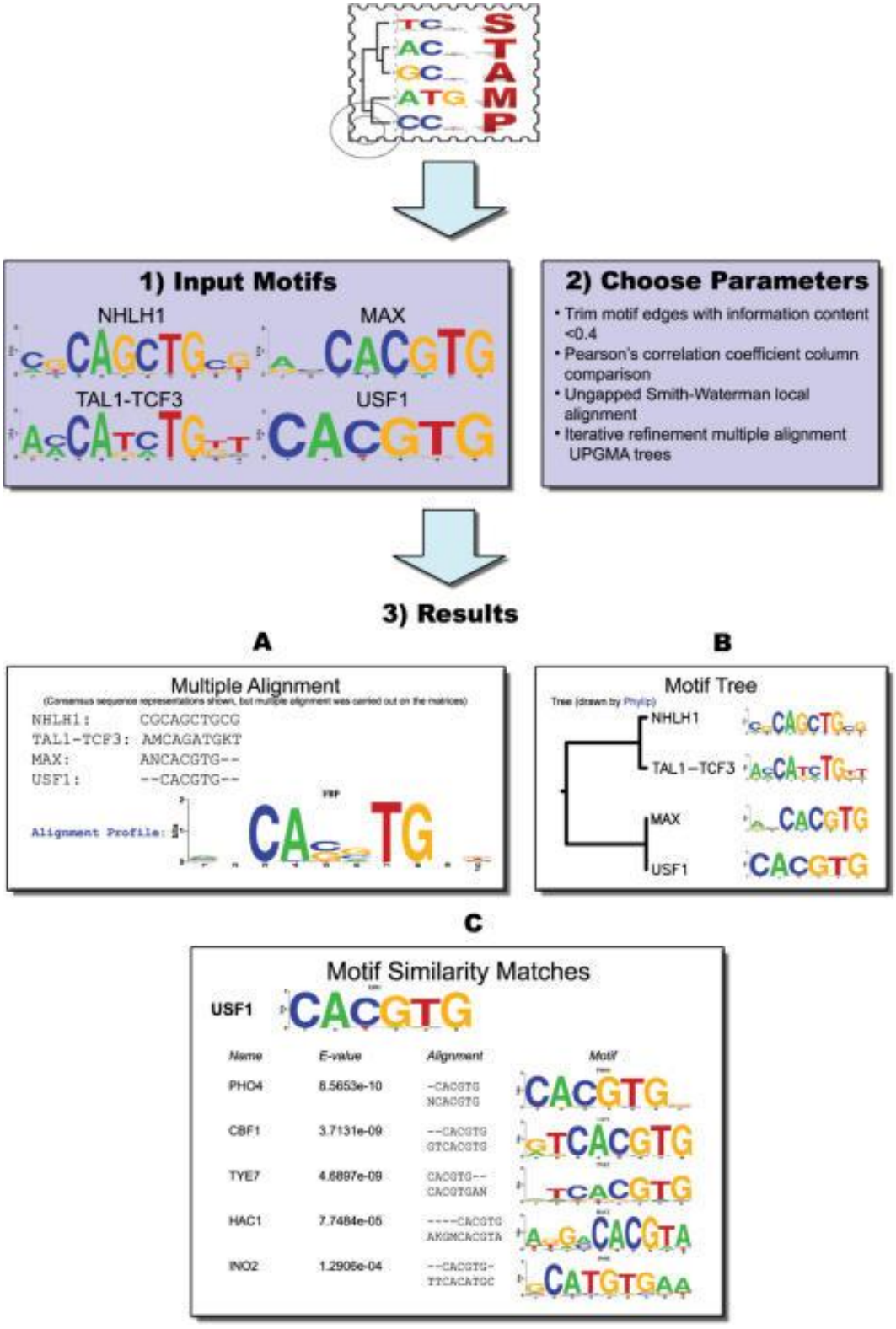


Figure 1.6: Analysis of four motifs using STAMP. (1) Position specific scoring matrices (represented as sequence logos) are given as input. (2) The user can either leave parameters to default (as shown in the figure) or choose different parameters for motif comparison. (3) After selecting a database for comparison and submitting a query, the results are divided in three parts. These parts correspond to: (A) a familial profile, which is the multiple alignment of all input sequences, (B) a similarity tree and (C) the most similar motifs to each of the input motifs present in the queried database (Mahony and Benos 2007).

STAMP produces similarity tree files in the Newick-format. Such files are viewable with the program *Molecular Evolutionary Genetics Analysis version 5* (MEGA5) (Tamura et al. 2011). The program offers several options for tree displaying, e.g. the branch length can be included in the tree visualization (see **Figure 1.7**). MEGA5 is freely available at http://www.megasoftware.net/.

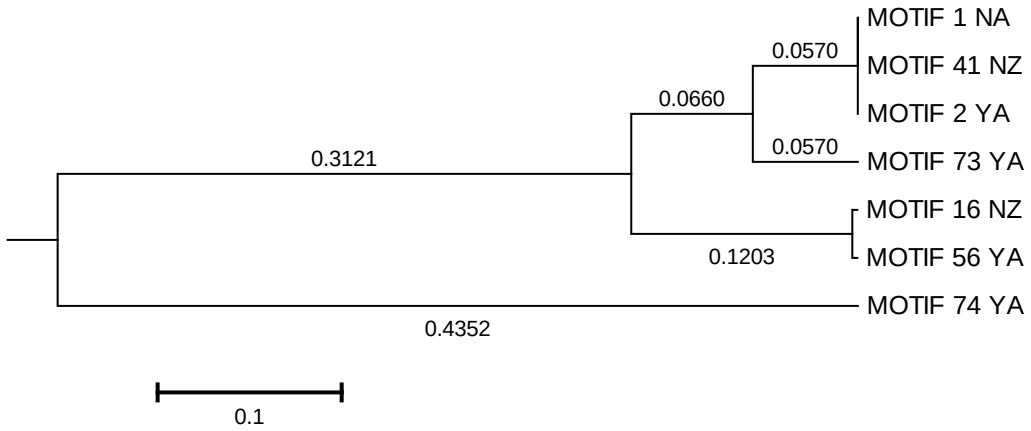


Figure 1.7: Example of a similarity tree viewed with the program MEGA5. Branch lengths can be displayed on the tree, lengths shorter than 0.008 were hidden for viewing.

1.7.3 Java

Programs developed in the course of the present study were written in the Java programming language. Java is an object oriented programing (OOP) language developed by James Gosling at Sun Microsystems (now owned by Oracle). Java was developed to be platform independent with the idea of "writing once, run everywhere" (Schildt 2011). Java uses classes to define objects, which are defined as data structures consisting of data fields and methods (Schildt 2011). In addition Java makes use of the three most important OOP principles: encapsulation, for controlling object access, inheritance, by which an object acquires properties of other objects and polymorphism, to confer an object the ability to belong to different types (Schildt 2011). Java relies on several libraries which contain classes for different purposes; built-in libraries contain classes with methods supporting Input/Output, string handling, networking and graphics, among others (Schildt 2011). These libraries, as well as external libraries, are available as application programming interfaces (APIs) (Schildt 2011).

1.7.4 Microsoft SQL

Gene expression and genomic data from the PathoPlant and AthaMap databases are stored and managed using the structured query language (SQL) MicrosoftSQL. MicrosoftSQL manages data in a relational database management system (RDMS) (Langenau 2001). In this relational model, a database is comprised of several two dimensional tables that have a relation between each other (Langenau 2001). These tables contain columns (attributes) and rows and represent real elements (Langenau 2001). Microsoft SQL uses an extended version of the language SQL called Transact-SQL .With the help of select statements information can be retrieved form the database in the form of row-sets, this information can come from one or different tables (Langenau 2001). The selection in a given statement can be filtered by adding certain conditions that the results should met in order to be retrieved (Langenau 2001). In addition, mathematical operations can be carried out on data from the tables which returns numerical values as results (Langenau 2001).

1.8 Goals of this study

Motif finding programs like MEME predict a large number of motifs. Trying to experimentally validate these large number of motif predictions is very difficult. Thus, there is an obvious need to further assess these motifs in order to determine which ones have a higher probability of being functional. Although some approaches have been reported to solve this problem (Bussemaker et al. 2001, Caselle et al. 2002), the focus in solutions for plants is still very limited. In order to fill this gap, the major goal of the present study is the identification of novel CREs in promoters of co-regulated *Arabidopsis thaliana* genes by means of newly developed bioinformatics methods. To accomplish this, a new tool was developed to *in silico* evaluate the probability that a sequence is a functional CRE responsive to different biotic and abiotic stresses. For this purpose, the tool correlates genome-wide sequence occurrences in promoters with *Arabidopsis thaliana* microarray expression data from the PathoPlant database. Furthermore, in order to facilitate element detection, new methods were implemented to assess the stress-specificity of a predicted CRE.

Given that transcription factors often bind to CREs in a cooperative manner, new methods were developed for the prediction of combinatorial CREs. Position-specific scoring matrices (PSSMs) were used to represent CREs. Promoter occurrences of PSSMs combinations were correlated with *Arabidopsis thaliana* gene expression data in order to find synergistic and putatively functional combinatorial CREs. The program also allowed the finding of combinatorial elements with and without characteristic spatial constraints.

After proof-of-concept, public-access tools were and will be released to the plant science community for the prediction of putatively functional CREs responsive to different abiotic and biotic stresses.

2 Methods

2.1 *Arabidopsis thaliana* genome and expression data

In this study several *Arabidopsis thaliana* genome-wide analyses were performed. For this purpose, *Arabidopsis thaliana* genomic data was retrieved from the Arabidopsis Information Resource (TAIR) (Swarbreck et al. 2008). Two widely used databases throughout this work were PathoPlant (Bülow et al. 2007) and AthaMap (Bülow et al. 2009). At the beginning of this study, these databases were using TAIR release 7 sequence and annotation data and for that reason this version was employed to perform all analysis. TAIR release 7 raw data is freely available for download at TAIR (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR7_genome_release/). These data are available in XML format. Using a Perl script (Lorenz Bülow, personal communication) all genome data, i.e. sequence and annotation data, was parsed and then saved into text files. The TAIR release 7 contains 31762 annotated genes.

Stress-related microarray expression data from the PathoPlant database was used to perform different calculations for the prediction of CREs. A complete list of all used data is given in **Table 7.1** (page 136). cDNA microarray experiments had previously been imported to the PathoPlant database as described in (Bülow et al. 2007). Further publicly available *Affymetrix* experiments (ATH1 and 8k chips, see array type column in **Table 7.1**) had been downloaded from TAIR and NASCAarrays, such data had been normalized and imported into PathoPlant (Bülow et al. 2007). In this study, an internal version of PathoPlant was used that also contains Zinc-related *Affymetrix* (ATH1) experiments provided by the lab of Ute Krämer.

2.2 Motif prediction with existing software

Prediction of overrepresented motifs in up-regulated gene promoters was carried out with the program MEME (Bailey et al. 2009). As described in **Chapter 1.7.1**, use of this program requires the specification of input sequences. Promoter sequences used as input sequences were extracted as described in **Chapter 2.2.1**. All motif finding parameters used with MEME are described in detail in **Chapter 2.2.2.**

2.2.1 Promoter sequences of co-regulated genes

Promoter sequences of up-regulated genes upon stresses shown in **Table 2.1** were used as input sequences for motif-finding with MEME. Such genes were identified using an in-house developed query tool (Lorenz Bülow, personal communication). The tool identifies genes which display an induction factor higher than a selected threshold for up to 6 different stresses. It builds SQL statements that retrieve *Arabidopsis thaliana* Gene Identifiers (AGIs) and promoter sequences for these genes from the PathoPlant and AthaMap databases. In order to do so, a stress or a combination of stresses from the PathoPlant database is first selected and then an induction factor threshold is defined. With such an information, a SQL statement is built, which will retrieve the promoter sequences and the AGIs of genes with an induction factor $\geq$ than the selected threshold. The length of input sequences has been shown to be critical for motif-finding algorithm performance (Hu et al. 2005). In addition, lengths longer than 500bp had been shown to decrease algorithm performance notably (Hu et al. 2005). For this reason the promoter length to be extracted was set to 500 bp.

Table 2.1: Stresses from the PathoPlant database used for identification of up-regulated genes.

Stress	**Time Point(s) or Concentration**
Chitooctaose	1hr
EF-Tu	30min, 1hr
Flg22 (P. syringae)	1hr, 2hr
Pb-oversupplied leaves	25ppm, 50ppm
Pb-oversupplied roots	25ppm, 50ppm
Zn-deficient roots	
Zn-deficient shoots	
Zn-oversupplied roots	2hr, 8hr
Zn-oversupplied shoots	8hr
Zn-resupplied roots vs. deficient Zn	2hr
Zn-resupplied roots vs. suffficient Zn	2hr
Zn-resupplied shoots vs. deficient Zn	8hr
Zn-resupplied shoots vs. suffficient Zn	8hr

As explained in **Chapter 1.7.1**, it has been shown that submitting more than 40 sequences does not improve motif finding by MEME. The number of sequences can be reduced by selecting a higher induction factor threshold. Therefore, different queries were performed in order to retrieve sets of genes containing about 40 sequences.

Since motif finding can also be improved when the number of input sequences is further reduced, further gene sets with sizes of around 20 and 15 sequences were extracted. The 40 most strongly up-regulated genes were identified by increasing the gene induction factor threshold until the desired number of genes was retrieved. Consequently, the induction factor threshold was increased until about 20 and 15 genes sets were also obtained. Promoters of those genes sets were used as input sequences for MEME.

2.2.2 MEME motif-finding parameters

Search for conserved motifs in promoters of up-regulated genes was carried out with MEME. The program takes input DNA sequences, extracted as described in the last chapter, and outputs overrepresented motifs. The number of output motifs can be adjusted within MEME. Version 4.3.0 of the program was downloaded from http://meme.sdsc.edu/meme/meme-download.html and locally installed on a Linux OS.

Table 2.2: Parameters used for motif finding with MEME. The number of input sequences (n) influences the number of minimum and maximum sites a motif can have within input promoters.

Motif Distribution	Minimum Sites	Maximum Sites	Minimum Width	Maximum Width
Any Number of Repetitions	Sqrt(n)	Min(5*n,50)	8	10
Zero or One Repetition	Sqrt(n)	n	8	10

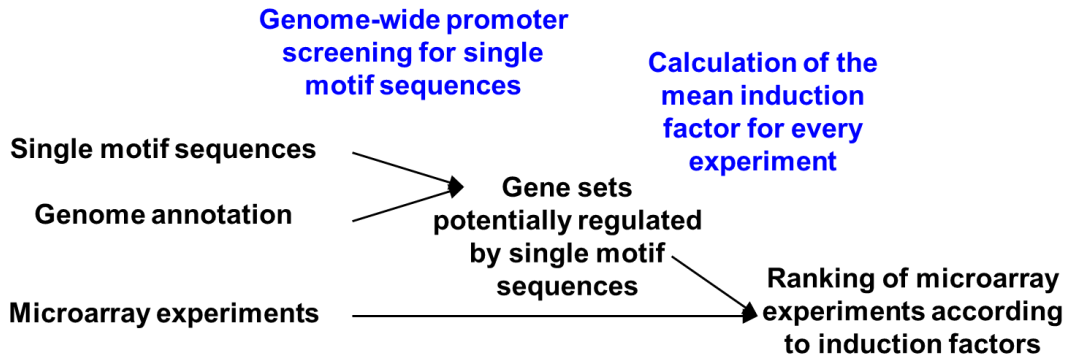
Several parameters have to be defined when using MEME to find motifs (see **Chapter 1.7.1**). **Table 2.2** summarizes all used parameters for motif finding with MEME: motif distribution, maximum number of motifs, minimum and maximum number of sites, minimum and maximum motif width. The motif distribution parameter is the type of occurrence distribution that single motifs have within the promoter sets (see **Chapter 1.7.1**). Two different distributions were selected for this parameter; *Zero or One Repetion per Sequence* and *Any Number of Repetitions per Sequence*. MEME also finds the optimal minimum number of sites for each motif within a user-established limit. This limit depends on the type of distribution selected and the number of input

sequences used. It is defined, for the number of minimum sites as *sqrt(n)*, *n* being the total number of input sequences. The maximum number of motif sites when the motif distribution is *Any Number of Repetitions per Sequence* is given by *min(5*n,50)*, and just *n*, when the distribution is *Zero or One Repetion per Sequence.* The optimum motif width is also selected by MEME within some user-defined limits. The minimum width limit was set to 8 and the maximum limit to 10.

Using these parameters, MEME identifies a very high number of overrepresented motifs among which also unspecific ones may occur. Because of this, predicted motifs with the MEME software were bioinformatically assessed for functionality with a newly developed tool described in **Chapter 2.3** and applying this tool, putatively functional motifs were identified as stated in **Chapter 2.5.**

2.3 *In silico* expression analysis method

In order to bioinformatically assess the functionality of predicted motifs, a new tool was developed. It is an *in silico* approach to validate sequences as CREs putatively responsive to different biotic and abiotic stresses. The tool uses microarray expression data from the PathoPlant database to calculate induction factor mean values of gene sets that contain a motif sequence within their promoters. The statistical significance of the average mean expression values is assessed by calculating a *p-value*. Such information is used to evaluate the probability of a given motif sequence to be a putatively functional CRE. **Figure 2.1** summarizes every step of the *in silico* expression analysis. The tool was programmed in Java1.6. The following chapters will describe in detail how this analysis is performed.



Sequences in genes showing relatively high induction factors under relevant conditions are potentially functional and specific.

Figure 2.1: Diagram showing every step of the *in silico* expression analysis. The tool uses single motif sequences and genome annotation data to perform a genome-wide promoter screening for single sequences. This results in gene sets potentially regulated by motif sequences. These sets are used to calculate mean induction factors from the PathoPlant database, which finally results in a ranking of experiments according to induction factors.

2.3.1 Genome-wide identification of promoters with single motif sequences

The *in silico* expression analysis starts with a genome wide promoter screening for motif sequence occurrences. The screening is summarized in **Figure 2.2**. For this purpose, all *Arabidopsis* gene promoters were extracted from the TAIR genome data files (see **Chapter 2.1**) as described next. Using information from the AthaMap database, the TSS, if known, otherwise the ATG site of all *Arabidopsis* genes was determined as the gene start position in order to extract the 500bp region upstream of this position. As used within the previous MEME analyses, the promoter of each gene was defined as the 500bp region upstream of the TSS or ATG position. Promoter sequences were stored in FASTA format files containing AGIs and the corresponding DNA sequences. These files are now accessed by the *in silico* tool in order to find exact matches of motif sequences in sense and antisense orientations within the gene promoters. Once a match was found, the corresponding AGI was stored in a list that resulted in a set of genes that contain a given motif sequence within their promoters (see **Figure 2.2**).

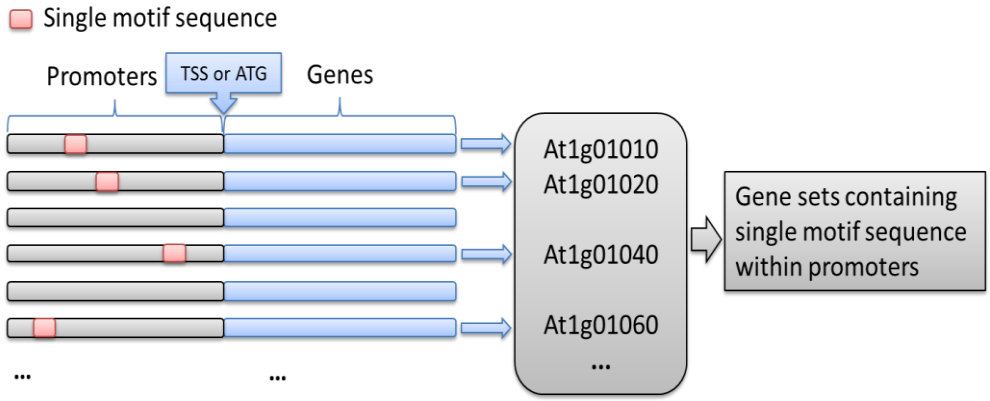


Figure 2.2: Promoter screening performed in the *in silico* expression analysis. Matches of single motif sequences (red boxes) are searched within gene promoters. This identifies gene sets that contain a single motif sequence within their promoters.

2.3.2 Mean induction factor calculation of gene sets

Once genes containing motif sequences within promoters were identified, they were used to calculate mean induction factors using microarray expression data for each one of the 155 experiments stored in the PathoPlant database. By building an SQL statement in which the gene sets AGIs are submitted as statement *conditions*, the corresponding gene induction factors were retrieved and mean induction factors for every single microarray experiment were calculated. The average mean expression *(Avg)* of a gene set *(w)* under a stress *(s)* is given by the geometrical mean of the induction factors:

$$Avg(w,s) = e^{\frac{\sum_{i=1}^{n} ln(F)}{n}} \qquad (1)$$

where

$$F = \begin{cases} \frac{1}{|fc|}, & \text{if } fc < 0 \\ fc\ , & \text{otherwise} \end{cases}$$

and *fc* denotes the induction factors `FOLD_CHANGE` value of a given gene under stress *s*. **Equation (1)** is applied for each microarray experiment and in this way, expression data for the gene sets under different stresses is retrieved. For comparability among the different experiments, these values were normalized

according to overall expression values of each microarray experiment. For this purpose, **Equation (1)** was also used to calculate the overall means of all genes for each of the different stresses. These values constitute normalization factors and were stored in a table that the *in silico* tool accesses in order to normalize each calculated mean value. The normalized values (*NAvg)* under stress *s* are given by

$$NAvg(s) = \frac{Avg(w)}{Avg(r)} \qquad (2)$$

where *Avg(w)* denotes the average mean expression for a gene set *w* under stress *s* and *Avg(r)* the average mean expression of all genes under stress *s*. The normalization values are shown on **Table 7.2**. After normalization, results were ordered according to mean induction factor values which resulted in a ranking list of experiments.

2.3.3 t-test statistics

Statistical significance of the average mean expression values calculated for the different gene sets was assessed by means of a *p-value* calculation with a t-test. The *p-value* is the probability that allows *null hypothesis* rejection in favor of an *alternate hypothesis* in order to assess if a given result is statistically significant or not. The *null hypothesis* was defined as: the average mean expression values measured are not significantly different than the overall average expression values and the *alternate hypothesis* as: the average mean expression values are significantly different than the overall average expression. The *p-value* is the smallest significance level at which one can reject the *null hypothesis*. In other words, the *p-value* gives the probability that the calculated mean value is not significantly different from the overall gene expression. A *p-value* was calculated for each measured average gene expression. This was done by implementing a student's t-test to assess the difference between the gene sets average expression means and the overall expression means under a given stress. For this purpose, the *Apache commons mathematics library* API version 2.0 was used. A *jar* library file that can be used to access all java classes contained within the API is available at http://archive.apache.org/dist/commons/math/source/ prior free license acquisition.

Following data are needed for *p-value* calculation from both, a given gene set and for all genes present in a microarray chip: average mean expression (*mean*), variance of the individual induction factors (*var*) and number of induction factors (*n*) used to calculate expression. These data were extracted from the PathoPlant database server each time a new calculation was performed. The class `TTestImpl` contained in the *Apache mathematics* API was then used to calculate the *p-value* with the method `homoscedasticTTest`. This method accepts *mean*, *var* and *n* as parameters in order to return the *p-value* as *observed significance* associated with a one-tailed unpaired t-test. These data was added to the output created by the *in silico* expression analysis tool, and in that way it was possible to determine the significance value of a calculated average mean expression value for a given stress.

2.3.4 Parameters for *cis*-regulatory element selection

Once mean induction factors (see **Chapter 2.3.2**) and *p-values* (see **Chapter 2.3.3**) were calculated, putatively responsive sequences to a given stress were identified. For this purpose a java tool was written that was able to identify CREs with several user-defined threshold parameters from the *in silico* expression analysis results.

There are three main criteria for CRE selection: *p-value*, number of genes containing motif sequences within promoters and ranking position according to induction factors. One very important criterion is the p-value (see **Chapter 2.3.3**). This value represents the significance of an average induction factor for a given stress and serves as a statistical measure and takes into account the mean induction factor for a given stress, the number of data used to calculate that value and the variance of that data (see **Chapter 2.3.3**). The p-value was always determined for all stresses and a threshold *p-value* was applied for the stress of interest, i.e. the stress a CRE is expected to be responsive to. Another important criterion is the number of genes containing a sequence within promoters. This value is always calculated with a genome-wide promoter analysis (see **Chapter 2.3.1**). Hence, a minimum value of genes containing a putative CRE within promoters was also defined. The results from the *in silico* expression analysis are ranked according to induction factors and significance values. This ranking also serves as an indicator of the most probable stress condition

associated with a CRE, given that it is possible to observe if there are other stresses with more significant p-values. Therefore, the position of a stress of interest within this rank was used as a selection criterion. In addition, a threshold for the average induction factor value was defined. To identify putatively functional sequences, all of these criteria were set as thresholds within the developed tool with the p-value being the most important one. The tool accesses output files generated by the *in silico* expression analysis and scans every sequence for values $\geq$ (or $\leq$ in case of p-values) than those set as thresholds. The sequences meeting the criteria are extracted and displayed by the tool.

2.4 *In silico* expression analysis validation

In order to establish criteria for CRE selection with the above mentioned tool and for validation of the newly developed *in silico* expression analysis method, known and novel synthetic CREs responsive to abiotic and biotic stresses were analyzed as controls. Sequences of known CREs were extracted from different publications and used as input data for the *in silico* expression analysis. The *in silico* expression analysis output was assessed for expected stress responsiveness. The sequences, the respective experimentally determined responsiveness and the corresponding literature source are shown in **Table 2.3**.

Table 2.3: Known CREs used as input sequences for *in silico* expression analysis validation

Sequence	**Responsiveness**	**Source**
TACCGACAT	Drought, low temperature and high-salt stress	(Yamaguchi-Shinozaki and Shinozaki 1994)
AGTTGACTAA	Plant defense mechanisms	(Ciolkowski et al. 2008)
ATGTCGACAT	Zinc deficiency	(Assunção et al. 2010)
ACGTCATAGA	Salicylic acid	(Johnson et al. 2003)

The novel synthetic CREs (synCREs) come from an experimental Chromatin Immunoprecipitation (ChiP) (Mario Roccaro personal communication) approach as described next. SynCREs were also used as input sequences for the *in silico* expression analysis and the information was used to assess if an enrichment of responsive elements was observed.

The chip method discovers new specific *cis*-acting elements from a library of synthetic elements by implementing a novel screening method. First, parsley protoplasts were transformed with constructs containing putative synCREs from a synthetic library, a minimal promoter and the *luc* reporter gene. Such protoplasts were tested for responsiveness upon elicitor treatment with Pep-25 derived from the fungus *Phytophthora sojae.* After adding the elicitor to the suspension culture, all actively transcribed gene fragments with their corresponding *cis*-elements are pulled-down. This is done first by performing a chemical cross-link of phosphorylated *Ser-5* residues of RNA polymerases II (RNApol II) carboxy terminal domains specifically bound by a phosphorylated *Ser*-5-RNApol II antibody. Once the chemical cross-link is performed, DNA-RNApol II complexes are extracted, sonicated and immunoprecipitated to identify actively transcribed gene fragments. Such fragments are then amplified via PCR, thus allowing specific synCRE enrichment. The synCREs were transformed again into parsley protoplasts and up to 3 rounds of enrichment (elicitor stimulation, chemical cross-linking, chromatin extraction, sonication, immunoprecipitation, PCR) were carried out. Enriched as well as non-enriched controls were subsequently sequenced in order to identify synthetic CREs responsive to fungal stress. **Figure 2.3** summarizes how sets were generated with the experimental approach (Mario Roccaro personal communication).

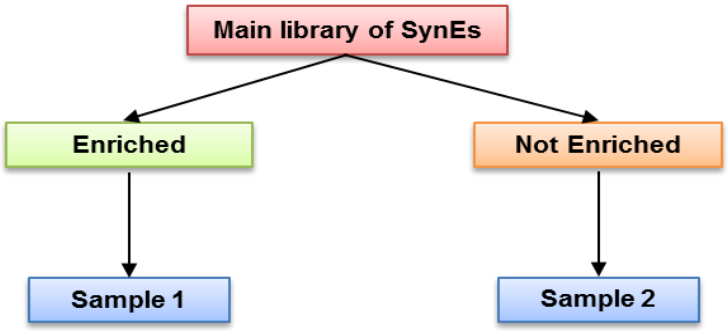


Figure 2.3: Diagram showing how sets were produced in the immunoprecipitation experiment. Synthetic elements (SynEs) of a main library were used for rounds of enrichment with Pep-25 elicitor from *Phytophthora sojae* and a sample containing enriched elements was produced. Another set of non-enriched elements were produced as a control.

Files containing sequences of the elements identified following the experimental approach were used as input sequences for the *in silico* expression analysis. The files correspond to samples 1 and 2, i.e. enriched and control, of **Figure 2.3** and they

contained variable 12 nucleotides long synthetic sequences, followed by a number indicating the frequency of that element in the sample. The 12N variable sequence was extracted and used as input sequence for the *in silico* expression analysis. Using the results from the analysis, overall expression values for all synthetic sequences from enriched and not enriched samples were calculated. In order to detect overall enrichment of responsive synthetic sequences in sample 1 compared to control sample 2, expression values for all elements in sample 1 were averaged by applying **Equation (1)** (see **Chapter 2.3.2**) to calculate the overall geometrical mean expression for each stress. The same calculation was performed using the output of sample 2. By calculating the ratio between sample 1 and sample 2 overall geometrical mean expression values for each stress, overall enrichment of synthetic sequences under specific stresses could be determined. A high calculated ratio for a given stress indicates that synthetic sequences responsive to this stress were enriched, whereas a low ratio indicates the opposite.

This analysis will enable identification of specific stresses for which the synthetic sequences are responsive, will help to refine the selection criteria of the *in silico* expression analysis, and finally serve as a proof-of-concept of the method if the stresses identified are related to fungal and other pathogens. For better comparison and for visualization, the ratios were organized in a ranking list from the highest to the lowest and they were used to construct Excel graphs that facilitated to assess whether the expected fungal or biotic responsiveness was observed. In addition, a p-value was calculated by comparing the average means calculated for sample 1 with the average means of sample 2. For this purpose, a t-test was performed where the averaged expression value for a given stress type in sample 1 was compared with the averaged expression value for that stress type in sample 2. Furthermore, it was tested whether the frequency of repetitions of a synthetic element has an effect on the overall expression values. For that purpose, the most frequent 10% and 5% of the elements in each sample (sample 1 and 2) were extracted. Again, overall average mean values were calculated by applying **Equation (1)** to each stress. The ratios between the samples and corresponding *p-values* were calculated as well. With these ratio values, Excel graphs were generated as described above in order to detect the specific stress

conditions for enrichment of synthetic *cis*-elements among the most frequent 10% and 5% of the elements in samples.

2.5 Analysis of MEME predicted motifs

Motifs were predicted using the program MEME as explained in **Chapter 2.2.** Such motifs were comprised of single sequences which in turn were used to test their putative functionality with the newly developed *in silico* expression analysis test (see **Chapter 2.3** for a detailed description on the analysis). By applying specific selection criteria, the output of the *in silico* expression analysis was now used to identify putatively functional CREs. Novelty and similarities among these predicted CREs were also assessed.

MEME predicts motifs which are comprised of single sequences. Following the methods described in **Chapter 2.3.4**, single motif sequences were identified as putatively functional by now applying specific selection criteria. 4 selection parameters were defined and are given in **Table 2.4**. These are minimum number of genes containing the CRE within promoters, minimum ranking position, minimum average induction and a maximum p-value, which corresponds to the significance of the average induction. These selection parameters were defined on the basis of validation experiments using known and novel synthetic CREs, as described in **Chapter 2.4.** However, since information is available only for the single sequences that comprise a motif and not for the motif itself as an entity, another selection parameter was introduced in order to select putatively functional CRE motifs entities. A motif was defined as putatively functional if at least one sequence met the selection criteria and at least 40% of the single sequences comprising the motif displayed significant p-values in respect to a given stress. By applying all these selection criteria, the putatively functional CRE motifs were identified.

Table 2.4: Parameters used to find putatively functional CREs from MEME motif sequences.

Stresses	**Promoters Containing CRE ≥**	**Ranking Position ≤**	**Average induction >**	**p-value ≤**
Chitooctaose, EF-Tu, Flg22 Pb-oversupplied, Zn-deficient, -oversupplied, Zn-resupplied	10	5	1.0	0.001

Novelty and similarity of the predicted motifs were assessed with the STAMP web server (Mahony and Benos 2007) and the program MEGA5 (Tamura et al. 2011) MEGA5 release #5110426 was downloaded from http://www.megasoftware.net and locally installed. STAMP allows querying motifs against databases in order to determine motif novelty (see **Chapter 1.7.2**). For this purpose, sequences comprising predicted motifs were extracted in FASTA format and they were used as input data for STAMP. Motif querying was performed against three plant databases of known transcription factor binding sites and CREs, AthaMap, Agris and PLACE, and search parameters were left to STAMP defaults. Motif similarity to known elements is reflected by low *e-values* obtained for each motif query and was used to determine novelty of the predicted motifs.

STAMP also produces a similarity tree that clusters the motifs into groups of related motifs. The tree format generated by STAMP is the *Newick*-format tree, which is viewable with the program MEGA5. The tree shows motifs that are grouped into clusters according to their similarities. Each branch holding a motif cluster has a particular distance that can also be shown with MEGA5. This branch distance was displayed within all similarity trees in order to define groups of related motifs. The branch length cut-off value was set to 0.008, which is even below the value used by the developers of STAMP to determine clustered motifs (Mahony et al. 2007). Motif diversity assessment was performed by grouping very similar motifs into clusters that were defined according to the branch distance cut-off value.

2.6 Selection of *cis*-regulatory elements with specificity and similarity information

A new approach was developed as an alternative to MEME for the prediction of input sequences to use in the *in silico* expression analysis. As CREs are typically 8 to 10 nucleotides long, most functional regulatory elements should be represented by a 10mer sequence. Therefore a set with all possible 10mer sequence combinations was generated. The sequences within the set were used as input sequences for the *in silico* expression analysis. This means that for every DNA 10mer, an *in silico* expression analysis was performed. The output of this analysis was used to find sets of CREs that are responsive to individual stresses. For this purpose, a new Java tool called *cis*-regulatory element finder (CREF) was developed to select from a single stress and certain search criteria CREs responsive to the selected stress (see **Figure 2.4**).

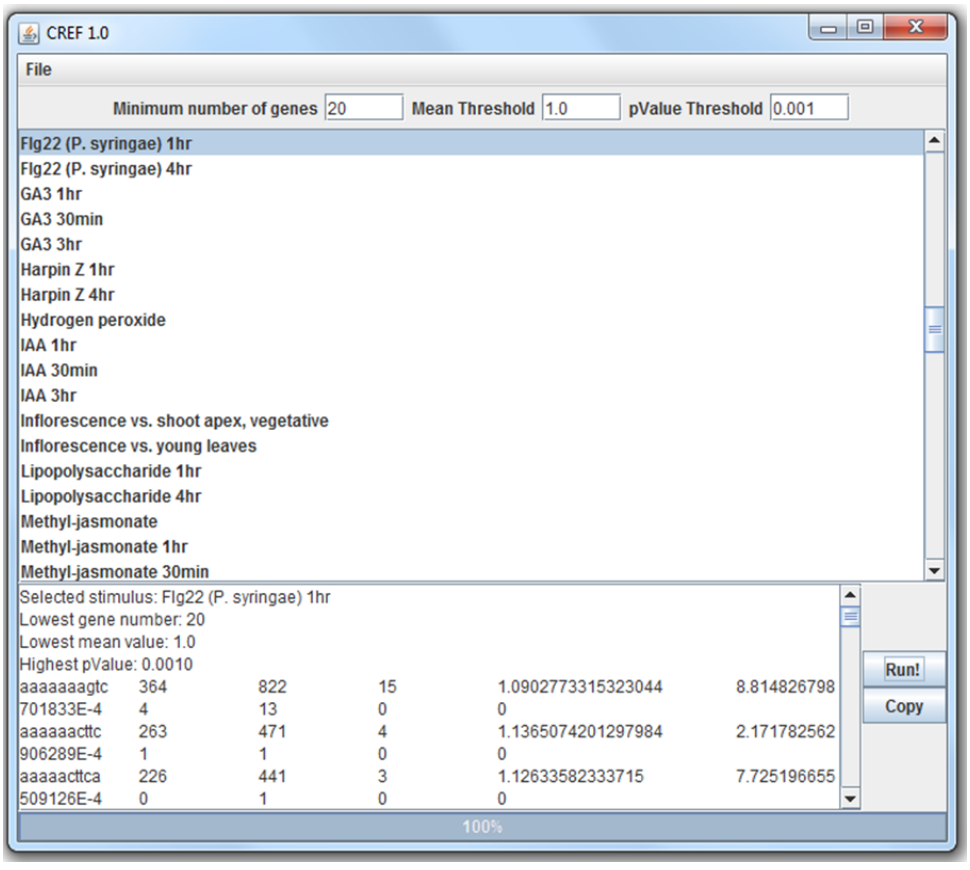


Figure 2.4: Screenshot of the tool *cis*-regulatory element finder (CREF). The tool allows the selection of the following parameter in order to find CREs: stress to which CREs should be responsive to, minimum genes number containing CRE within promoters, lowest mean induction factor of such genes upon selected stress and p-value calculated for that mean induction factor.

Sets of CREs putatively responsive to Flg22 and Drought stresses were predicted since they are important representatives for a biotic and for an abiotic stress. shows the steps followed for CRE sets prediction. Such new predictions could be experimentally tested.

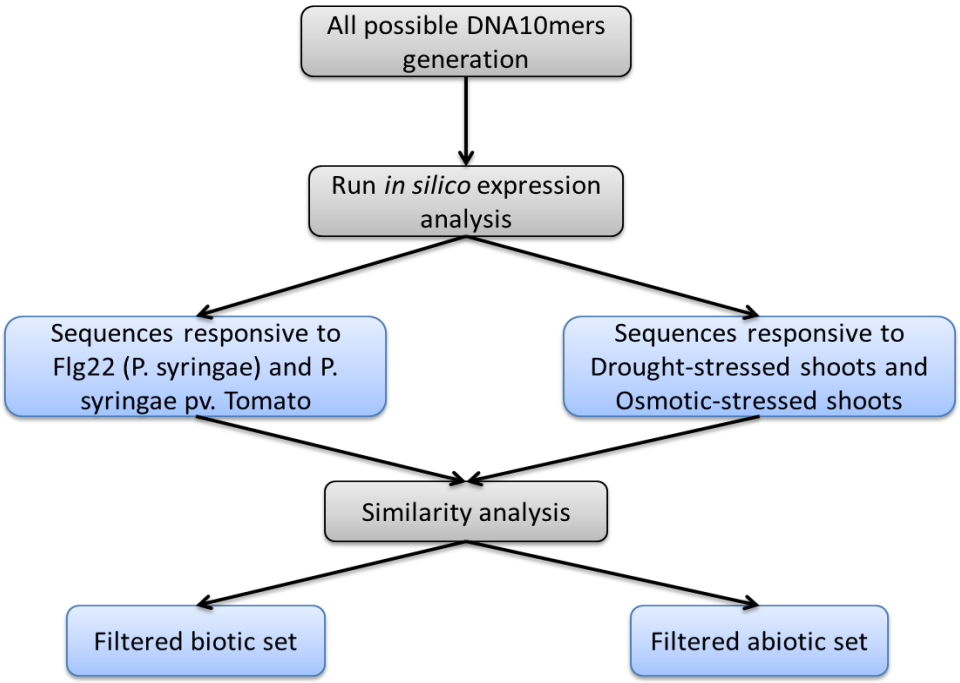


Figure 2.5: Diagram showing the pipeline followed to find highly specific CRE sets. The process started by generating all possible DNA10mers which were then used as input sequences for the *in silico* expression analysis. After that, sets of sequences putatively responsive to Flg22 (P. syringae) and P. syringae pv. Tomato and to Drought-stressed shoots and Osmotic-stressed shoots were predicted. With such sequences a similarity analysis was performed that allowed the identification of putatively specific CREs.

From the set of all possible DNA10mers used as input for the *in silico* expression analysis, two CRE sets, one putatively responsive to Flg22 (*P. syringae*) and *P. syringae* pv. tomato, and another set to Drought-stressed shoots and Osmotic-stressed shoots were identified. The CREs of these two sets were predicted as described next. First, using the tool CREF (see **Figure 2.4**), putatively functional sequences were identified. The selection criteria for CREF were: a minimum of 20 genes should contain the sequence within promoters, genes should be up-regulated (i.e. mean induction factor above 1.0) upon selected stress and the p-value should be $\leq$ 0.001. Then, the specificity of each sequence was assessed by identifying if other stresses also displayed significant p-values. The type (biotic, abiotic, fungal and other) of such stresses was determined and the number of stresses of each type showing significant p-values was identified.

This information about specificity was used to further select sequences from the Flg22 (*P. syringae*) and *P. syringae* pv. tomato set which displayed no abiotic stresses with significant p-values. Similarly, sequences from the Drought-stressed shoots and Osmotic-stressed shoots set showing no significant values for biotic stresses were selected. This resulted in one biotic and one abiotic set. Finally, in order to introduce another level of specificity both sets where compared in order to identify sequences showing no similarities to abiotic sequences, for biotic responsive sequences and sequences showing no similarities to biotic sequences, for abiotic ones. This was achieved by generating a similarity tree with the STAMP web server and by identifying clusters containing only abiotic or biotic sequences. Flg22 and Drought responsive sequences were identified in such clusters to generate two putatively functional sets. Similarity trees were constructed for each set using the STAMP webserver (see **Chapter 1.7.2**), where it was possible to determine if the predicted sequences show similarities to any previously reported CRE.

2.6.1 Abiotic stresses

In the course of the analyses it turned out that abiotic stresses seem to nearly always be present within the ranked lists described before. In order to have a closer look at the response to the abiotic stresses salt, osmotic, cold and drought, all possible CREs responsive to these abiotic stresses were identified. For this purpose, the tool *cis*-regulatory element finder (CREF) (see last chapter) was used to identify CREs responsive to Cold, Osmotic, Salt and Drought stresses at different time points (see **Table 2.5**).

Table 2.5: Abiotic stresses analyzed.

Stress	**Time points in PathoPlant**
Cold-stressed roots and shoots	0.5hr, 1hr, 3hr, 6hr, 12hr and 24hr
Drought-stressed roots and shoots	0.25hr, 1hr, 3hr, 6hr, 12hr and 24hr
Osmotic-stressed roots and shoots	0.5hr, 1hr, 3hr, 6hr, 12hr and 24hr
Salt-stressed roots and shoots	0.5hr, 1hr, 3hr, 6hr, 12hr and 24hr

Finally in order to visualize the sequence similarities and the number of CREs among the abiotic stresses, area-proportional venn diagrams were generated. With such diagrams it was possible to compare the number of CREs and visualize how many

overlapping CREs are present among the abiotic stresses compared. This facilitates a detailed investigation of the abiotic response.

2.7 Pathway crosstalks

A new crosstalk analysis method was developed as described in this chapter. The main idea of the analysis is that crosstalk among different stresses occurs and genes are responsive to several stresses due to convergences in the signaling pathways upstream of the transcription factors (see **Chapter 1.3**). This means that transcription factors can be associated to several stresses and thus, a CRE can also be responsive to several stresses. In this work, gene sets containing CREs within promoters were identified as being putatively responsive to a certain stress, such genes can also be expected to be responsive to other stresses. The *in silico* expression analysis (see **Chapter 2.3**) was developed to predict CRE responsive to a certain stress. It is also possible to assess responsiveness to all other PathoPlant stresses with this tool allowing identification of possible signaling pathway convergences with CREs as starting points for the analysis. This approach was used as described in the next chapters to assess if a given CRE set is putatively responsive to several stresses.

2.7.1 Predicted motifs

Crosstalk analyses were performed for CREs responsive to the stresses shown in **Table 2.1**. Such analyses yielded information about which other possible stresses the CREs can be responsive to. In order to calculate such information, a java tool was developed. In this crosstalk analysis, the *in silico* expression analysis output of all elements in a putatively functional regulatory elements set is used to calculate overall expression values of all stresses in the whole set. This is done by calculating the overall geometric mean values, i.e. averaging all expression values of a stress *s* observed in a set *c,* as given by

$$OE(c,s) = e^{\frac{\sum_{i=1}^{n} \ln(Avg)}{n}} \quad (3)$$

where *Avg* denotes calculated average expression under stress s of a gene set containing a single CRE within promoters and *n* denotes the total number of observed *Avg* in *c*.

By applying **Equation (3)** to a given set of regulatory elements, a ranked list of overall expression values of all stresses stored in the PathoPlant database is obtained. All the elements in the set were initially predicted to be responsive to a stress *S*, and in order to assess which other stresses are the most similar to stress *S* (i.e. the stress the regulatory elements had initially been identified as putatively responsive to), a p-value was calculated in respect to the initial stress. The p-value was calculated as described in **Chapter 2.3.3** with some changes in the input parameters as described next**.** Following data were used for p-value calculation: overall average expression (*mean*), variance (*var*) and number of average means (*n*) used to calculate overall expression. For each p-value calculation the *mean*, *var* and *n* of the stress of interest and the same respective values of a given stress were used as input parameters for the `homoscedasticTTest` method defined in the class `TTestImpl` (see **Chapter 2.3.3**) in order to calculate a p-value. The resulting value serves as a similarity measure given that if the p-value is bigger than the threshold 0.001 it is not possible to reject the *null hypothesis.* This *null hypothesis* states that there is no significant difference between the expression of the stresses. The p-value is defined as the probability that the overall average expression values from the stress of interest are not significantly different from the overall average expression values of the stress being compared. In this way it was possible to determine if a regulatory element set is responsive to further stresses. Multiple stresses associated to such a set of regulatory elements indicate putative crosstalks between these stresses.

2.8 Combinatorial *cis*-regulatory elements

Combinatorial action of CREs was assessed with a newly developed pipeline. Several methods were implemented in order to find synergistic combinations of CRE. **Figure 2.6** summarizes steps followed in order to predict putatively functional combinatorial CREs.

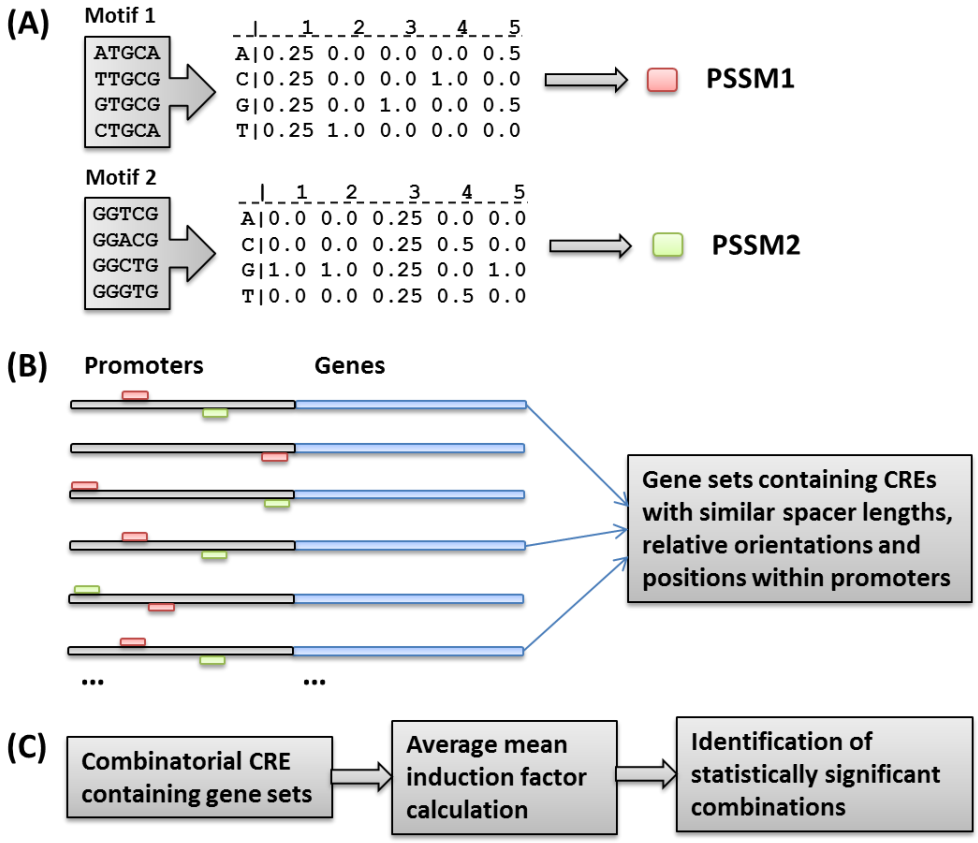


Figure 2.6: Pipeline to find combinatorial CREs. In a first step (**A**) PSSMs are build using sequences which comprise a motif. Such matrices are used for a genome-wide promoter screening (**B**) in order to find promoters containing combinations of both PSSMs. An algorithm is then used to find gene sets containing CREs with similar spacer lengths, relative orientations and positions within promoters (**B**). The position constraint can be changed in order to obtain combinatorial CREs with different positions. In a last step (**C**) identified gene sets are used to calculate average mean induction factors and to find statistically significant combinatorial CREs.

Position Specific Scoring Matrices (PSSMs) were used to represent CREs. Thus, finding of combinatorial CREs starts by building PSSMs using single motif sequences. For this purpose, overrepresented motifs predicted by MEME in the gene promoters of the 40 most up-regulated genes upon each analyzed stress, were used to generate PSSMs. Such matrices were constructed from the motif sequences by calculating the observed nucleotide frequencies at each position in the motif and by tabulating such values in a matrix. Defined formally, from a sequence set S forming a motif and with n sequences of length m, $s_1, \ldots, s_n$, where $s_{kj} = s_{k1}, \ldots, s_{km}$, and s_k denotes one of the nucleotide symbols (A, C, G, T) at position j in the sequence s, a PSSM M_{4xm} is made as

$$M_{ij} = \sum_{k}^{n} I_i(s_{kj}) \qquad \begin{matrix} i = A, C, G, T \\ j = 1, \dots, m \end{matrix} \qquad (3)$$

where

$$I_{(q)} = \begin{cases} 1 & if \quad i = q \\ 0 & otherwise \end{cases}$$

By applying **Equation (3)** to a motif (**Figure 2.7A** gray box) a PSSM is obtained, where the number of times a nucleotide is seen at each position is indicated in the motif matrix. These absolute frequencies were then used to calculate relative frequencies, which are defined as the nucleotide occurrences at a given position divided by the number of motif sequences (**Figure 2.7B**).

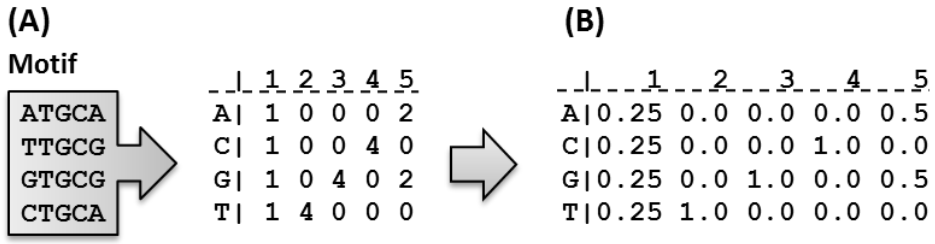


Figure 2.7: Matrix representation of a motif. The simplest PSSM of the motif shown in the gray box is shown on (**A**). Each position on such matrix denotes the number of times a nucleotide was seen at that position. Relative frequencies of a nucleotide at each position are shown on (**B**), where numbers represent the probability of seeing a nucleotide at a given position.

After PSSMs were constructed, they were used to perform genome-wide promoter screenings for CRE motifs. This was achieved by scanning gene promoters for the presence of similar sequences to the profile represented by the PSSM. A score was calculated in order to assess if both sequences were similar. Such scores were produced by running a window with the length of a given PSSM along the promoters and by summing the corresponding nucleotide coefficients at each position in the window. For a sequence window *w* with length *m* a score (based on a PSSM *M*) was defined as

$$m_w = \sum_{j=1}^{m} M_{wij} \qquad (4)$$

where $w = w1, \ldots, w_m$ and w_i denotes one of the nucleotide symbols (A, C, G, T). In this way scores were obtained for each position within promoter sequences. A score threshold for considering a sequence similar to the PSSM profile was defined as described next. The score of each sequence comprising a motif was calculated using **Equation (4)** and the matrix derived from the motif. The minimum score of those sequences was set as the threshold. The promoter screening is presented in **Figure 2.8**.

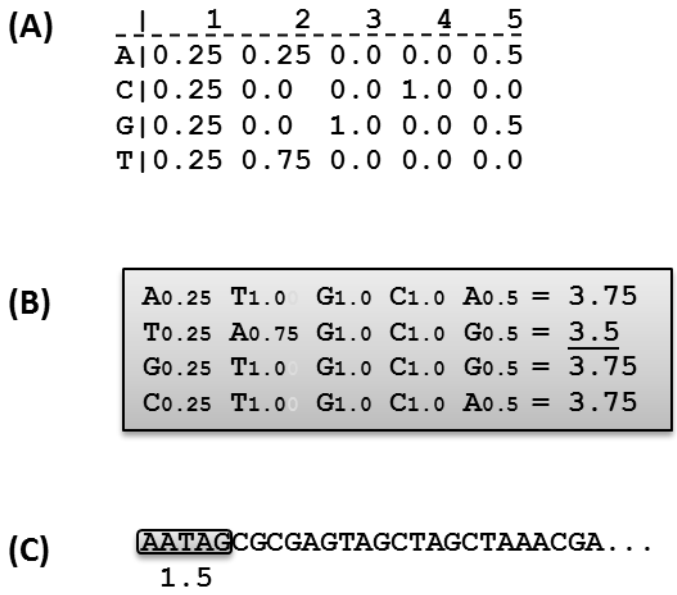


Figure 2.8: Promoter screening using PSSMs. Given a PSSM **(A)**, sequences from which the matrix was derived are used to calculate scores **(B)**. The lowest score is set as threshold for considering a sequence similar to the PSSM profile (3.5 in this example). Finally **(C)** a window (gray box) with the PSSM length is moved across promoters to produce scores at each sequence position. Thus, allowing determination of sequences similar to the PSSM profile.

Once positional information was gathered for each CRE motif using PSSMs, gene sets containing combinations of CRE were identified. Given that a goal of the present study was to develop a program that could also predict combinatorial elements with similar spatial constraints (spacer distances, motif order and motif orientation), an algorithm was developed and applied to determine which genes harbor CRE motifs with similar spatial constraints. The spacer in a combinatorial element was defined as the distance

from the end of a CRE motif to the start of another CRE motif (see **Figure 2.9**). Such distances were calculated for all predicted combinatorial elements. In order to identify combinatorial elements with similar spacer lengths, genes containing the same pair of motifs with a spacer length that was allowed to wobble ±5 nucleotides were identified. This wobble factor was chosen because a full DNA turn comprises 10 nucleotides. Furthermore, the algorithm can also identify combinatorial elements with the same motif order and the same relative orientations, which allowed the identification of combinatorial elements with similar spatial constraints (see **Figure 2.10**).

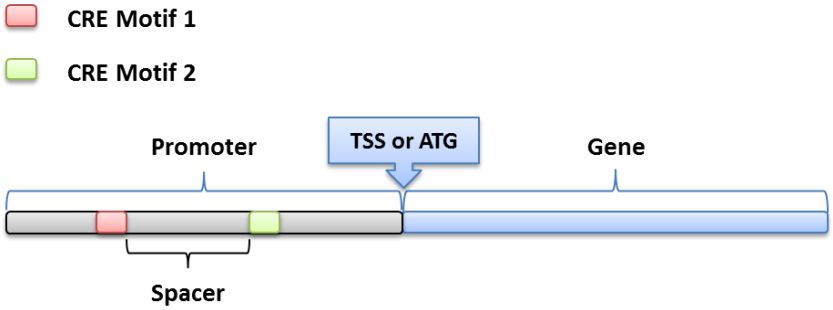


Figure 2.9: Spacer between two CRE motifs. A spacer was defined as the distance between the end of a CRE motif and the start of another CRE motif.

In order to determine the probability of the predicted combinations being a combinatorial CRE, average gene expression and statistical significances (by means of a p-value) were calculated similar to the *in silico* expression analysis using single CREs. The average expression values of gene sets containing combinatorial CREs within promoters were calculated as explained in **Chapter 2.3.2** which resulted in a ranking of microarray experiments according to induction factors. The statistical significance of the average expression values was determined as described in **Chapter 2.3.3**. Information gathered with the analysis was used to identify statistically significant combinatorial CREs (p-value $\leq$ 0.001 for the stress of interest) and elements showing the stress of interest in a ranking position $\leq$ 5. Thus, a combination of CREs was said to be functional when these conditions were complied. The predicted element sets were used to assess their similarities between each other and to determine if the predicted sets have characteristic spacer lengths and distances to the TSS.

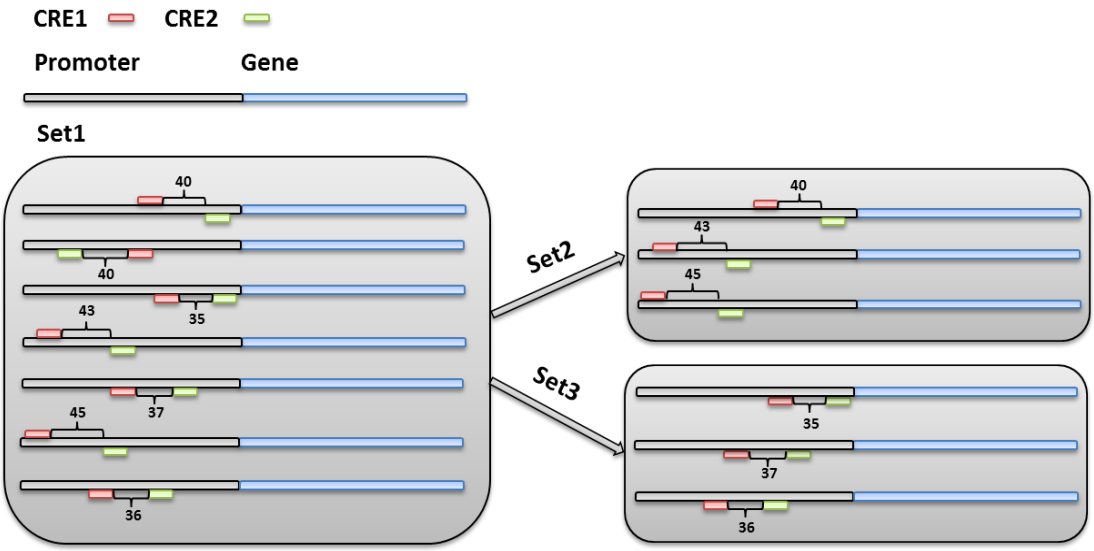


Figure 2.10: Finding of combinatorial CREs with similar spacer lengths, orientations and positions. First a set (Set1) of genes containing 2 CREs is identified. Next, new gene sets are determined (Set2 and Set3) where CREs: have: spacer length distances that differ ±5; have the same relative orientations; and have the same motif order within promoters. It was also possible to identify gene sets containing combinations of CREs with different motif order within promoters.

In order to visualize all gathered information for the putatively functional combinatorial CREs an internal web tool was developed (Yuri brill, personal communication). The tool was written in the PhP programming language. It produces a graphic where the promoters containing the combinatorial CREs are shown. The tool uses information in XML format to be able to draw representations of promoters and CREs. For this purpose code was implemented in the Java pipeline to store all positional information in XML format. An example of the graphics produced by the tool is shown in **Figure 2.11**.

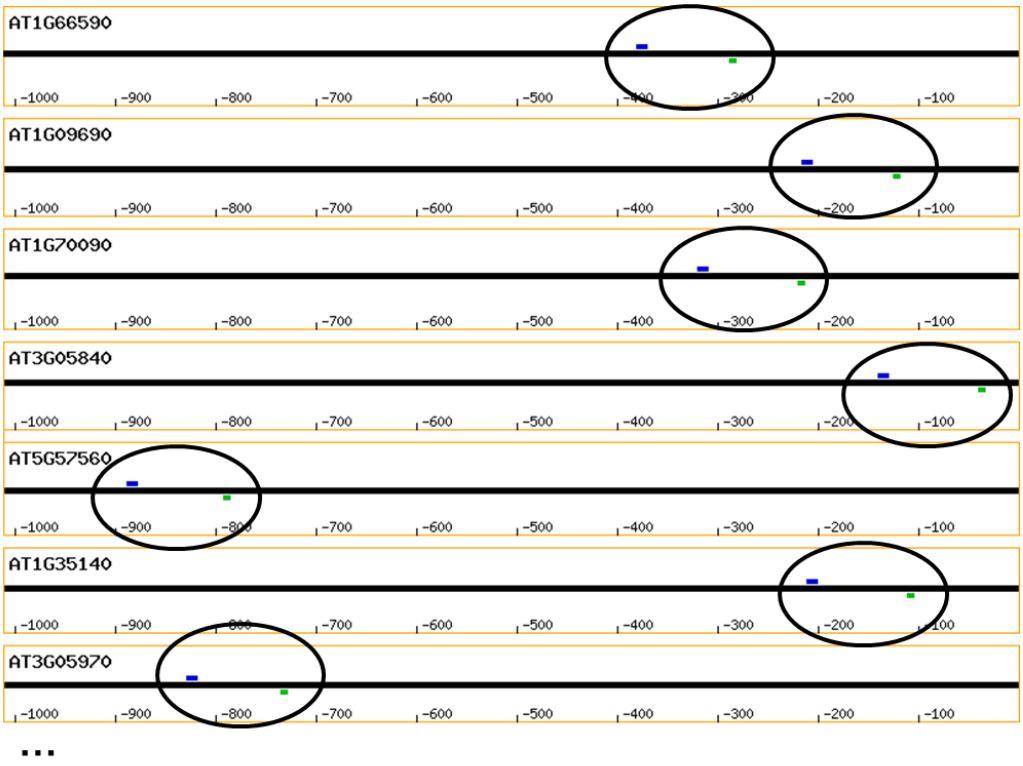


Figure 2.11: Output of combinatorial CRE visualization tool. Each yellow rectangle represents a gene promoter. CREs are shown as small blue and green rectangles and are highlighted in circles.

2.8.1 Similarity analysis

After having predicted sets of putatively functional combinatorial CREs, a measure of the similarities among the elements was needed. For this purpose the STAMP webserver and a newly developed java tool were used as described next. First, using STAMP, similarity tree files were produced using the motifs comprising combinatorial CREs. A java tool was developed that read such files in order to determine combinatorial CRE similarities. In the trees, motifs are clustered together according to their similarities. Branches linking motif clusters have a length that serves as a measure of motif similarity. Motifs within a cluster whose branch distance length was <0.008 were considered similar. Thus, it was defined that two combinatorial CREs were similar, if the elements forming the combination come from the same cluster (see **Figure 2.12**). For this purpose the java program determines to which cluster each single CRE belongs to and with that information it assess which combinatorial elements are formed with elements from the same cluster (see **Figure 2.12**).

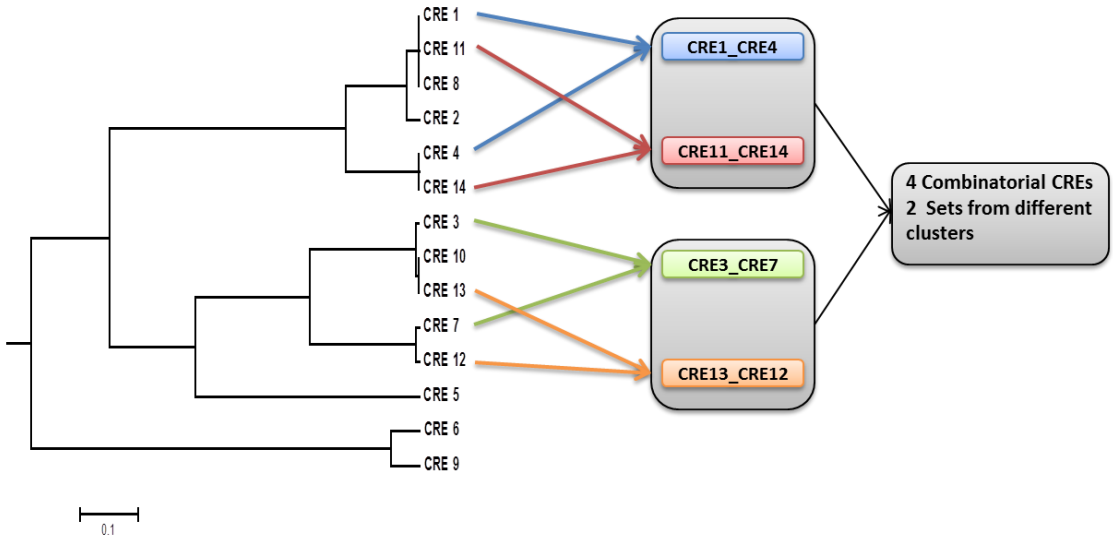


Figure 2.12: Similarities of combinatorial CREs. A similarity tree produced with STAMP and viewed with MEGA5 is used to assess how similar a set of combinatorial CREs (blue, red, green and orange boxes) are. Individual elements from a combinatorial CRE belong to different clusters in the similarity tree. A program assess to which cluster each CRE is grouped, in order to determine if more combinations exist from motifs of the same clusters. In that way it is possible to determine the number of different combinatorial CREs.

2.8.2 Spacer length analysis

Sets of combinatorial elements with spacer lengths constraints were predicted. In order to assess if there were characteristic lengths among the combinatorial elements in the different predicted sets, the analyses described in this chapter were carried out. The frequencies of observed spacer lengths were compared with random distributions to determine if such frequencies were higher than randomly expected. For this purpose, a background model was generated that shows how many elements are randomly expected with certain spacer lengths in a combinatorial element set. 18 different sets of *n* combinatorial CREs putatively responsive to biotic and abiotic stresses (see **Table 2.1** in page 24) were predicted and background models for each predicted set were produced. Such background models were generated with a newly developed program which generated sets of *n* random combinatorial elements. The program determined the number of elements in these random sets with spacer lengths within a range of 50bp, i.e. it determined how many random elements have spacers with a length of 0 to 50bp, 51 to 100bp, … , and 951 to 1000bp. For each of the 18 predicted sets, 10^5 random sets were generated and their values were averaged in order to obtain the theoretical number of expected combinatorial elements with

spacer lengths in ranges of 50bp. The calculation of the average expected elements at a given range *r*, was given by

$$r = \frac{\sum_{i=1}^{n} d}{n} \qquad (5)$$

where *n* is the number of random sets generated (10^5) and *d* represents the spacer length frequencies observed at range *r*. For example a background model for a set of 200 combinatorial elements calculated as explained before results in the model shown in **Figure 2.13**, where each column represents the average number of expected elements with a given spacer length.

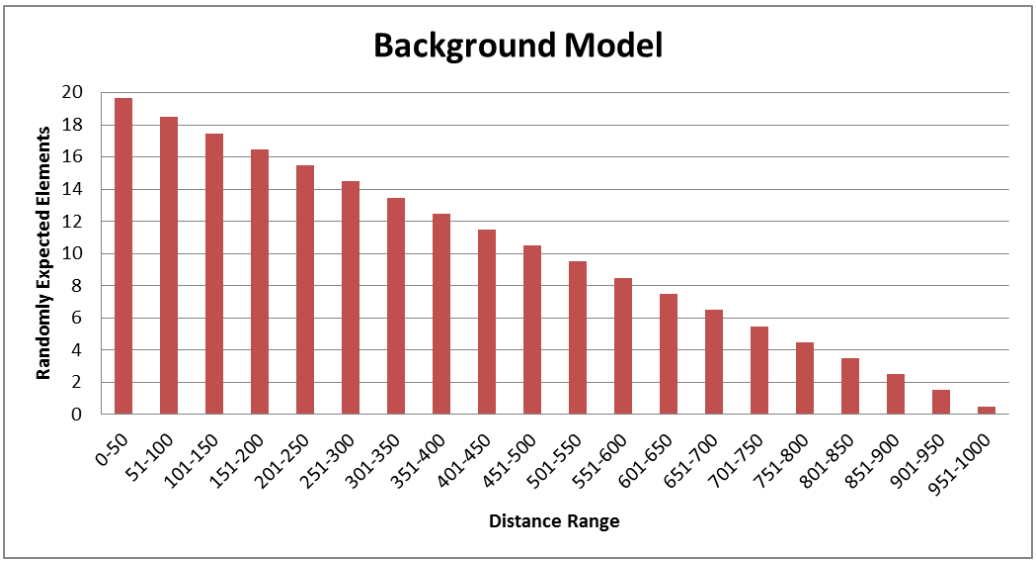


Figure 2.13: Background model of randomly expected elements for a set of 200 combinatorial CREs. The figure shows in a random distribution how many elements (*y* axis) are expected with a spacer length that lies in a range of 50bp (*x* axis).

In order to determine if there are significant and characteristic differences between observed spacer length frequencies and expected spacer length frequencies, the Poisson probability was calculated. It indicates the probability *P* that exactly *x* combinatorial elements occur given an average expected value *λ,* the Poisson distribution is given by

$$P(x, ë) = \frac{ë^x e^{ë}}{x!} \quad (6)$$

and it was calculated for each distance range and each background model.

2.8.3 Distance to the TSS

Another goal of the present study was to test if the predicted combinatorial elements in a set display characteristic distances to the TSS. For this purpose, distances to the TSS from the motifs forming combinatorial elements were calculated. For this analysis two different distances were defined (see **Figure 2.14**): the distance to the TSS from the nearest CRE motif and the distance from the farthest CRE motif forming a combinatorial element.

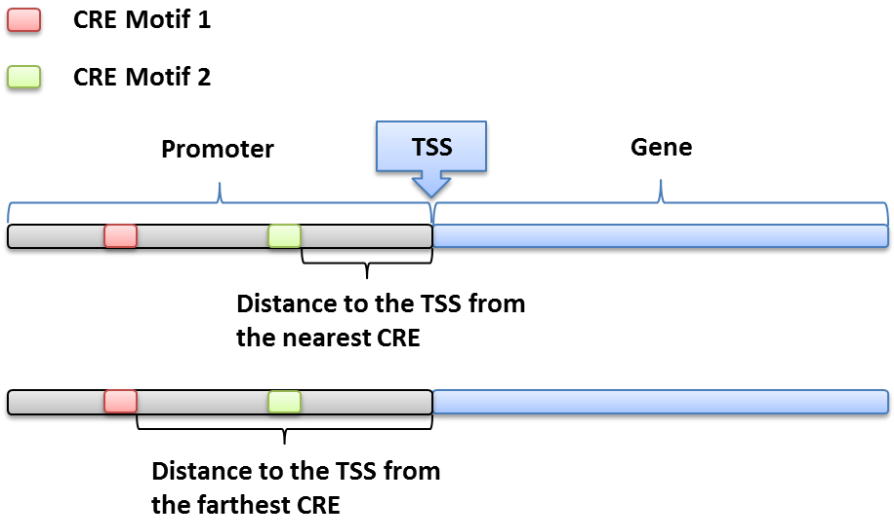


Figure 2.14: Combinatorial CRE distance to the TSS. Two different distances to the TSS were measured: the distance from the nearest CRE and the distance from the farthest CRE forming a combinatorial element.

18 different sets of combinatorial CREs putatively responsive to biotic and abiotic stresses (see **Table 2.1** in page 24) were predicted. For each of the combinatorial elements present in these sets, the distances to the TSS of the nearest and farthest CREs comprising the combinatorial elements were calculated. In a similar approach as the one explained in the last chapter, the frequencies of distances to the TSS with a length lying within a range of 50bp, i.e. 0-50bp, 51-100bp, … , and 951-1000bp were

calculated. In addition it was tested if these distances follow a random distribution. This was accomplished by generating two background models (one for the nearest and one for the farthest distances to the TSS see **Figure 2.14**) for each predicted combinatorial element set. Such models would allow the comparison of expected and observed distance frequencies and they were constructed as described next. Each of the 18 predicted combinatorial element sets had a different number *n* of elements. Thus, for each predicted set 10^5 sets with *n* random combinatorial elements were generated, their distances to the TSS were measured and the values were averaged using **Equation (5)** (see page 48), which resulted in theoretical expected numbers of distances to TSS. Thus, two background models were produced for each predicted set; one for the nearest and one for the farthest motifs to the TSS (see in **Figure 2.15** a background model for a set of 200 combinatorial elements). Finally, in order to test if the differences between observed and expected distances were significant, **Equation (6)** was applied.

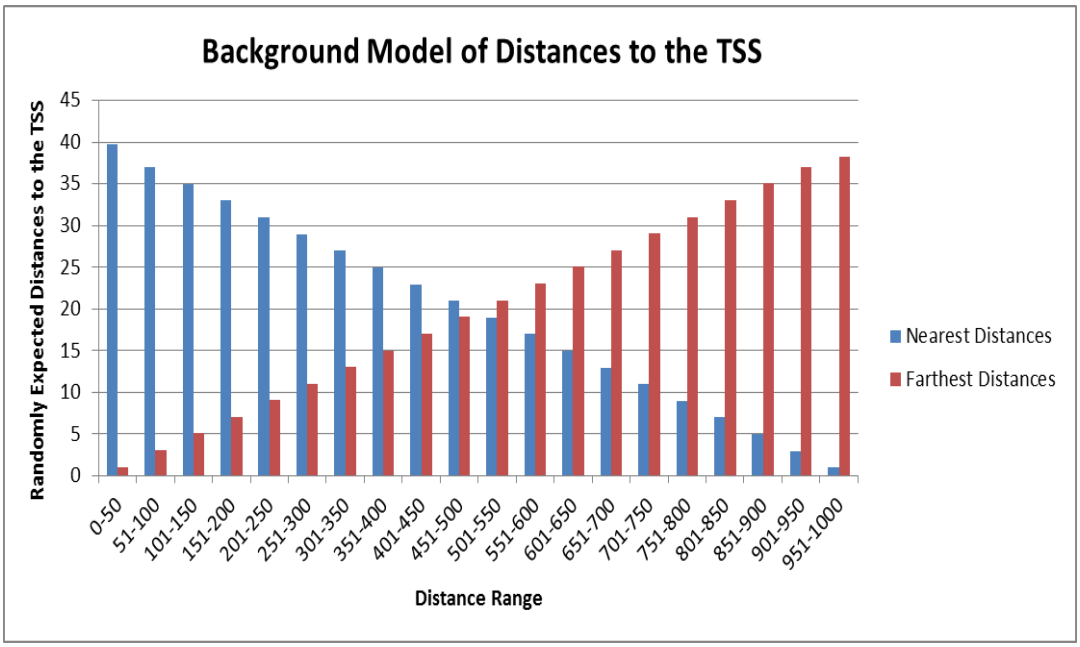


Figure 2.15: Background model of randomly expected distances to the TSS in a set of 200 combinatorial elements. The model shows in a random distribution of how many distances to the TSS (y axis) are expected with a length that lies in a range of 50bp (x axis) for the nearest (blue columns) and farthest (red columns) distances.

3 Results

3.1 *In silico* expression analysis to validate motif predictions

The main goal of the present study was to predict putatively functional *cis*-regulatory elements (CREs). For this purpose, the tool *in silico* expression analysis, described in **Chapter 2.3,** was developed. It is a bioinformatics approach to determine the probability of a given sequence to be a functional CRE. This tool was used in the present study to predict several putatively functional CREs. The tool was validated by assessing if the predicted responsiveness of certain known CREs from literature are consistent with the expected response (see **Chapter 3.1.1**). In addition, a whole set of pathogen-responsive synthetic CREs from a high-throughput experimental screening was also analyzed with the *in silico* expression analysis tool in order to assess it (see **Chapter 3.1.2**). After these validations, the tool was then used to predict novel putatively functional CREs responsive to biotic and abiotic stresses (see **Chapter 3.1.3**).

3.1.1 Known *cis*-regulatory elements as proof of concept

The basis for all CRE predictions of the present study was the *in silico* expression analysis tool, described in detail in **Chapter 2.3**. The tool uses short DNA sequences as input and correlates gene promoter occurrences of such sequences with microarray expression data from the PathoPlant database. This tool then needed to be validated in order to be confident about the predictions performed with it. For that purpose, known CREs were used as input sequences in the *in silico* expression analysis and the results were assessed to determine if the CREs showed the expected response. The information gathered with this analysis was also used to define selection criteria for novel CREs predicted with the tool. The results obtained for each known CRE used to validate the *in silico* expression analysis are described next.

The Drought Responsive Element (DRE) with the sequence TACCGACAT was reported by (Yamaguchi-Shinozaki and Shinozaki 1994) as being responsive to drought, low temperature and high salt stress. This sequence was used to perform an *in silico* expression analysis. A genome-wide promoter screening (see **Chapter 2.3.1**) identified

53 genes containing the sequence within promoters (500bp upstream of the TSS if known, otherwise the ATG site). Average expression values of these genes were also calculated as described in **Chapter 2.3.2** with all biotic and abiotic microarray experiments. Each expression value has a corresponding p-value which provides a way to assess its statistical significance (see **Chapter 2.3.3**). By assessing the output of the *in silico* expression analysis it was possible to observe that mainly abiotic stresses are associated to the DRE. Only cold, osmotic, salt and ABA stresses have a p-value <0.001 (see **Table 3.1**). The ABA responsiveness of the sequence is also expected given that the element is involved in ABA-associated response (Yamaguchi-Shinozaki and Shinozaki 1994). Salt and drought stresses are predicted with a p-value of 3.63E-04 and 4.77E-03 respectively, which indicates that the DRE can also be responsive to those stresses but mainly to cold stresses.

Table 3.1: Stresses showing p-values <0.001 after *in silico* expression analysis for sequence TACCGACAT.

Stress	**p-value**
Cold-stressed shoots 24hr	4.38E-21
Cold-stressed roots 24hr	4.90E-21
Cold-stressed roots 12hr	3.10E-17
Cold-stressed shoots 12hr	1.22E-15
Cold-stressed roots 6hr	8.65E-07
Osmotic-stressed roots 3hr	6.24E-05
Cold-stressed shoots 6hr	2.67E-04
ABA 3hr (10µM)	2.92E-04
Salt-stressed shoots 24hr	3.63E-04

Another known CRE evaluated with the *in silico* expression analysis was the WRKY binding site AGTTGACTAA (Ciolkowski et al. 2008). This sequence is known for being involved in plant defense mechanisms (Ciolkowski et al. 2008). The sequence is present in 84 gene promoters. Such genes show average induction factor values with significant p-values mainly for biotic stresses. Shown in **Table 3.2**, only three biotic stresses show a p-value <0.001. This shows, as expected, that the sequence could be involved in the regulation of biotic stresses.

Table 3.2: Stresses showing p-values <0.001 after *in silico* expression analysis for sequence AGTTGACTAA.

Stress	**p-value**
P. syringae pv. phaseolicola 6hpi	8.12E-06
TMV systemic leaves 14dpi	8.36E-04
P. syringae pv. tomato hrcC- 2hpi	8.67E-04

A newly reported zinc-deficiency responsive CRE with the sequence ATGTCGACAT (Assunção et al. 2010) was also analyzed. The internal version of PathoPlant contains microarray expression data for zinc deficiency (Pajonk, personal communication). Therefore it was possible to assess the responsiveness of the sequence upon this stress. The CRE element is a palindromic sequence and is present in 26 gene promoters. The *in silico* expression analysis results show that the stress *zinc-deficiency roots* is the most probable condition associated with this sequence (see **Table 3.3**). The genes containing the element show also significant values for further zinc-deficiency related stresses and some biotic stresses (see **Table 3.3**). Overall, the results strongly suggest that the element should be responsive to zinc deficiency mainly in roots, which accords with the expected response.

Table 3.3: Stresses showing p-values <0.001 after *in silico* expression analysis for sequence ATGTCGACAT.

Stress	**p-value**
Zn-deficient roots	5.53E-18
Zn-deficient shoots	2.11E-06
Chitin 6hr	4.23E-05
Chitin 3hr	3.29E-04
Zn-resupplied shoots 8hr vs. suffficient Zn	3.52E-04
Zn-resupplied roots 2hr vs. suffficient Zn	6.23E-04
P. infestans 24hpi	6.65E-04

Finally, a CRE responsive to salicylic acid was also analyzed. The sequence ACGTCATAGA, reported by (Johnson et al. 2003), is present in 14 gene promoters. The *in silico* expression analysis indicates that, in regard to expression, genes containing this sequence show the stress salicylic acid as the stress with the most significant p-value. This is also very interesting, as this is exactly the expected responsiveness.

Table 3.4: Stresses showing p-values <0.001 after *in silico* expression analysis for sequence ACGTCATAGA.

Stress	**p-value**
Salicylic acid	6.48E-7
P. infestans 24hpi	1.20E-04
Zn-deficient roots	3.26E-04

Together these results serve to validate the *in silico* expression analysis tool. All cases indicate that the expected responsiveness of each CRE is present. In addition, this data was used to define selection criteria of novel CREs. As explained in **Chapter 2.3.4**, there are three very important parameters that can be used to select CREs. Thresholds for such parameters were determined with the results shown above. The first parameter is the p-value. It was seen for all elements analyzed that the stress of interest displays a p-value <0.001, which makes this a reasonable value to be set as threshold for CRE selection. The second parameter is the number of genes containing the sequence within promoters. Known elements analyzed are present in at least 10 gene promoters and therefore that number was set as threshold. Another parameter is the position of the stress of interest, i.e. the stress the element is responsive to, within the *in silico* expression analysis ranking. From the observed results for the known CREs it was observed that stresses of interest lie within the top 5 stresses in the ranked results. Therefore it was defined that the stress of interest should also lay within the top 5 stresses in the *in silico* expression analysis results. Other stresses, apart from the stress that the CREs was expected to be responsive, were also observed with significant p-values. This is a possible indicator of stress crosstalks (see **Chapter 1.3**) which was analyzed in depth as described in **Chapter 3.2.**

3.1.2 Pathogen responsive synthetic *cis*-regulatory elements

As another validation approach for the *in silico* expression analysis, synthetic CREs (synCREs) from a high-throughput experimental approach (Mario Roccaro, personal communication) were analyzed. This experimental method aims to discover CREs responsive to fungal elicitation from a random library of synthetic elements by implementing a novel screening method driven by enrichment steps with the elicitor Pep-25 derived from the fungus *Phytophthora sojae*. Two sets of synCREs were isolated with this experimental approach: one set with 3096 elements after treatment

with Pep-25 (enriched set) and one untreated control set with 2801 elements. Such synCREs were used as input sequences for the *in silico* expression analysis, which also determines the stress-specific expression of the genes that contain the sequences within their promoters. For all individual stresses, the overall geometrical average expressions of all elements were calculated. This was done separately for both, the enriched and the control set (see **Chapter 2.4**). Using these means, both sets were compared by calculating the ratios of the expression values for each of the 155 stress conditions represented by microarray data, and these ratios were used to construct **Figure 3.1** in order to graphically assess if an expected fungal responsiveness of the enriched set was detectable. For simplicity the figure only identifies microarray data for fungal, biotic and other experiments. In the graph, a value above one means that overall expression for that stress is higher in the enriched sample than in the control sample, whereas a value below one means the opposite. These ratios are organized from the highest to the lowest and colors were assigned to differentiate the stresses on the graphs. Red columns represent fungal stresses, blue columns represent biotic stresses excluding fungal and yellow columns represent other stresses or conditions (abiotic, development, hormones and signal molecules). By evaluating **Figure 3.1** it is possible to observe that the biotic and fungal stresses have some of the highest ratios of expression values, indicating that there is a higher number of synCREs putatively responsive to these stresses in the enriched sample. On the other hand, stresses from the group others have the lowest expression values, which indicates that there are less synCREs in the enriched sample responsive to these other stresses. These results show that the synCREs from the enriched sample seem to be specific and responsive to fungal and biotic stresses.

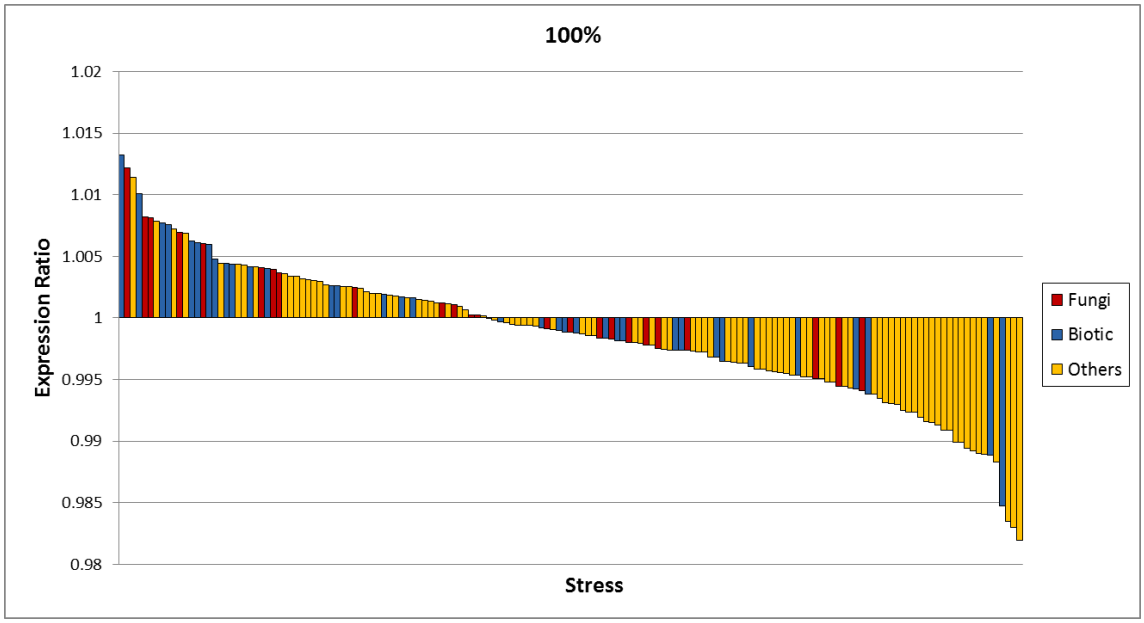


Figure 3.1: Expression values comparison between enriched and control samples. Each column represents the ratio of the overall mean calculated for a stress and the color of the column represents the stress type. Values above and below one indicate higher and lower expression values for a given stress in the enriched sample when compared with the control sample.

Since the synCREs isolated were determined using 454 high-throughput sequencing (Mario Roccaro, personal communication), each of the identified synCREs from the experimental screening was associated with a certain frequency within the samples. It was expected that synthetic elements with a high representation would also display an increased, i.e. a more specific and higher response towards fungal stresses. Therefore, the frequency effect of a synthetic element was tested (see **Chapter 2.4**). For that purpose, the most frequent 10% of the elements in each sample was extracted. With such elements the analyses described above were performed, i.e. the difference between their overall average means was determined. **Figure 3.2** displays the comparison between the most frequent 10% elements in the enriched and control samples. It shows a similar but stronger effect than the one observed in **Figure 3.1**, which is that biotic and fungal stresses clearly have the highest ratios of expression values and that stresses from the group “others” have the lowest expression values, which indicates that there is a higher number of synCREs putatively responsive to fungal and biotic stresses in the enriched sample.

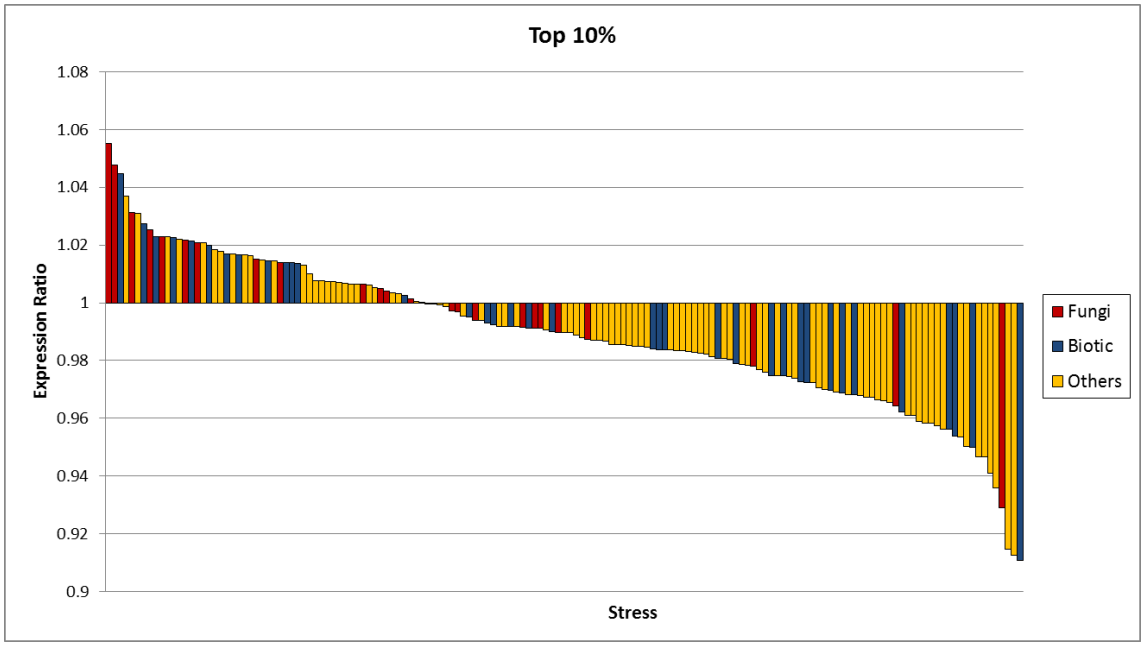


Figure 3.2: Expression values comparison between the most frequent 10% elements in the enriched and control samples. Each column represents the ratio of the overall mean calculated for a stress and the color of the column represents the stress type. Values above and below one indicate higher and lower expression values for a given stress in the enriched sample when compared with the control sample.

Even more stringently, the effect of the most frequent 5% elements within the sets from the enriched and control samples was also assessed. The ratios of the overall average means were determined to construct **Figure 3.3**. The elements in the enriched set now show an even clearer majority of fungal stresses with the highest overall expression ratios. Furthermore, stresses from the group others and the group biotic have even lower overall expression ratios than in the most frequent 10% samples. This means that the elements of the most frequent 5% enriched sample are more specific to fungal stresses than the other analyzed samples (10% and total). Taken together, these results nicely serve as a validation of the *in silico* expression analysis but also of the experimental method, since the expected response, i.e. the enriched elements show a specific fungal responsiveness, was clearly demonstrated.

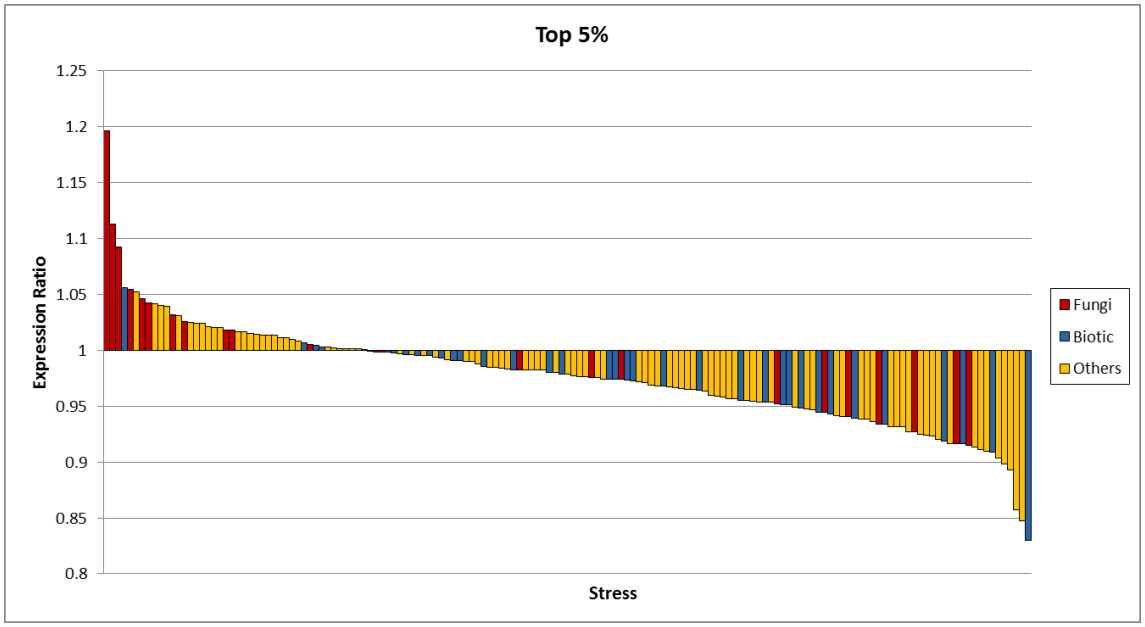


Figure 3.3: Expression values comparison between top 5% enriched and not enriched synthetic elements. Each column represents the ratio of the overall mean calculated for a stress and the color of the column represents the stress type. Values above and below one indicate higher and lower expression values for a given stress in the enriched sample when compared with the not enriched sample.

3.1.3 Identification of novel putatively functional *cis*-regulatory elements

Having tested the functionality of the *in silico* expression analysis as a prediction tool (see last two chapters), novel CREs responsive to selected biotic and abiotic stresses were predicted. As described in **Chapter 2.2**, overrepresented motifs were identified in promoters of up-regulated genes by using the program MEME. Such promoters were extracted as described in **Chapter 2.2.1** and used as input data for motif prediction with MEME (see electronic appendix **E_3.1.3/Genes_and_promoters_for_MEME_analysis** for the predicted genes in each stress condition and the induction factors used to predict them). The program yielded 6700 possible motifs (see **Table 3.5**) for all analyzed stresses. Sequences comprising such motifs were used as input data for the *in silico* expression analysis. This allowed the identification of 1014 motifs which met the selection criteria defined in **Chapter 2.5**. All motifs are shown in the electronic appendix **E_3.1.3/Predicted_Motifs.** The redundancy among predicted motifs was assessed by constructing similarity trees with the STAMP web server (see **Chapter 2.5**). Such trees were generated for each set of motifs responsive to the analyzed stresses.

Table 3.5: Number of motifs predicted after MEME, *in silico* and Stamp analyses.

	Number of motifs		**Number of Clusters**
Stress	**MEME Analysis**	***In silico* expression analysis**	**STAMP Analysis**
Zn-Deficiency	880	58	17
Zn-Oversupply	1060	18	10
Zn-Deficiency vs resupplied	1400	52	21
Pb-Oversupply	1480	637	150
Flg22	800	64	17
Chitooctaose	320	18	10
EF-TU	760	167	36
Total	6700	1014	261

Figure 3.4 shows the similarity tree for elements predicted to be responsive to Flg22. Each cluster contains motifs which are expected to be very similar. For Flg22, the 64 motifs predicted by the *in silico* expression analysis were grouped into 17 similar clusters, each with a different number of motifs (see **Figure 3.4**). Similarity trees of the remaining predicted motifs are shown in **Chapter 7.3.** 58 motifs were predicted to be responsive to Zn-Deficiency and such motifs were further grouped into 17 similar clusters. From the 1400 motifs predicted by MEME for the stress Zn-Deficiency vs. resupplied, 52 were predicted to be putatively functional by the *in silico* expression analysis and such motifs formed 21 different clusters. The number of motifs predicted to be responsive to Pb-Oversupply was 637 and they clustered into 150 different groups, which is very high in comparison with the other analyzed stresses. Another stress showing a high number of predicted elements is EF-Tu (167), although the similarity of such motifs seems to be high, which can be seen by the much smaller number of different clusters observed (36). On the other hand, the number of motifs responsive to Chitooctaose and Zn-oversupply (18) was low in comparison to other stresses. Both motif sets were grouped together into 10 different clusters.

With STAMP is also possible to assess the similarity between the predicted motifs and known functional CREs. For this purpose motifs were compared with known elements

from the databases Agris, AthaMap and Place (see **Chapter 1.6**). Such analyses were performed for each predicted motif set (see **E_3.1.3/Predicted_Motifs**). **Table 3.6** lists some examples of predicted motifs putatively responsive to Flg22. The first motif is similar to the known pathogen-responsive CRE WBOXATNPR1 (Yu et al. 2001). As can be seen in the sequence logo, both sequences are very similar which leads to a low e-value (a measure of the similarity between both motifs) and therefore indicates high similarity. **Figure 3.4** highlights all the motifs displaying similarities to known pathogen-responsive CREs (see motif names with red circles). There are also a high number of motifs which do not show significant similarities with pathogen associated CREs. Cluster 1 for example contains motifs very similar to the CRE CCTCGTGTCTCGMGH3 (Ulmasov et al. 1995) from the Place database. The element conferred inducibility towards the hormone auxin in tobacco seedlings (Ulmasov et al. 1995). One motif of cluster 1 is shown in the second entry of **Table 3.6**. Motifs in cluster 14 contain the "T-box" named TBOXATGAPB (Chan et al. 2001) (see the third entry of **Table 3.6** for an example), an element that has been shown to play a role in the transcription of light-activated genes (Chan et al. 2001). Thus, since all predicted motifs were extracted with same selection criteria, it is expected that the previously unreported motifs found in the sets also display the expected response.

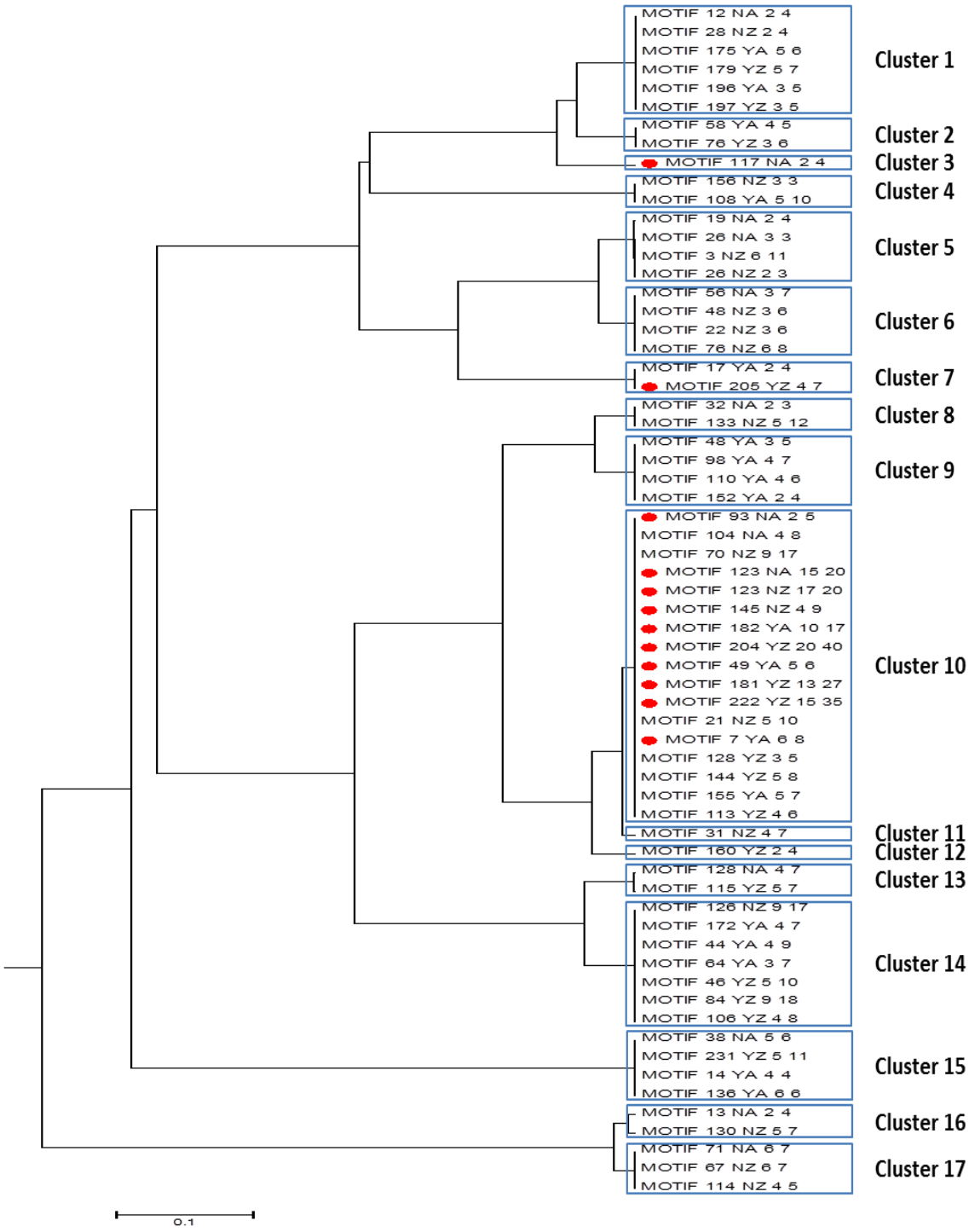


Figure 3.4: Similarity tree of Flg22 predicted CREs. Motifs with similarities to known pathogen responsive CREs are marked with a red circle. Very similar clusters are highlighted with blue boxes for better visibility.

Table 3.6: Examples of predicted motifs putatively responsive to Flg22 with the most similar motif from the place database.

Familial profile sequence logo	Motif, Source, name and e-value of similar CRE	Sequence logo of similar CRE
	Motif 125 NA, Place WBOXATNPR1 7.0204e-07	
	Motif 197 YZ, Place CCTCGTGTCTCGMGH3 1.2226e-09	
	Motif 64 YA, Place TBOXATGAPB 1.1561e-07	

Other previously reported and unreported putatively functional motifs were also predicted for other stresses. **Table 3.7** shows examples of predicted motifs putatively responsive to Chitooctaose (see electronic appendix **E_3.1.3/Predicted_Motifs/ Chitooctaose/** for a complete motif list). Similar motifs to the abscisic acid responsive element ABREMOTIFIIIOSRAB16B reported by (Ono et al. 1996) were predicted to be responsive to Chitooctaose in the present study. Abscisic acid has been show to play crucial roles in *Arabidopsis thaliana* responses towards plant pathogens (Fan et al. 2009). An example of a novel motif is also shown in **Table 3.7**. Another predicted motif was found to have some similarities with the CRE SURECOREATSULTR11 reported by (Maruyama-Nakashita et al. 2005). The CRE was reported to be involved in *Arabidopsis thaliana* responses to sulfur (Maruyama-Nakashita et al. 2005), but it has not been linked with pathogen responses.

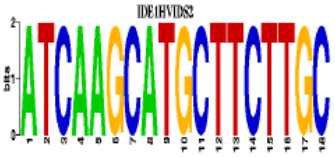
Table 3.7: Examples of predicted motifs putatively responsive to Chitooctaose with the most similar motif from the place database.

Familial profile sequence logo	Motif, Source, name and e-value of similar CRE	Sequence logo of similar CRE
	Motif 30 NA, Place ABREMOTIFIIIOSRAB16B 7.5029e-07	
	Motif 6 NA, Place SURECOREATSULTR11 1.6260e-05	

As mentioned before, a similar mixture of reported and unreported motifs was predicted for the other analyzed stresses. **Table 3.8** summarizes some examples of predicted motifs for different stresses. Motifs with similarities with known pathogen

responsive CREs were predicted for the stress EF-Tu (see electronic appendix **E_3.1.3/Predicted_Motifs/EF-Tu/** for a complete motif list). A motif similar to the CRE WBBOXPCWRKY1 reported by (Eulgem et al. 1999) to be involved in plant pathogen defenses was predicted to be responsive to EF-Tu. An example of a novel motif (MOTIF 55 NA) also predicted to be responsive to EF-Tu is shown in **Table 3.8**. The motif displays similarities to the CRE ANAERO2CONSENSUS which is involved in the fermentative pathway (Mohanty et al. 2005) but has not been reported to be involved in plant pathogen responses. A large number of motifs were predicted for the metal stress Pb-Oversupply (see electronic appendix **E_3.1.3/Predicted_Motifs/Pb-Oversupply/** for a complete motif list). Two examples are shown in **Table 3.8**. The Motif 64 YA has a high similarity with the metal-related element IDE1HVIDS2, which was reported to regulate iron deficiency (Kobayashi et al. 2003). The Motif 348 YZ is similar to the abiotic-stress related CRE DRE1COREZMRAB17, involved in water stress regulation (Busk et al. 1997). Among the motifs predicted to be responsive to Zn-Deficiency, a majority of elements with sequences similar to the known zinc-responsive sequence ATGTCGACAT (Assunção et al. 2010) (also shown in **Table 2.3**) were predicted (see **E_3.1.3/Predicted_Motifs/Zn-Deficiency/** for a complete motif list). The element seems to be crucial for *Arabidopsis thaliana* responses to Zn-Deficiency (Assunção et al. 2010) and shows a high similarity to the CRE CRTDREHVCBF2 reported to regulate gene regulation under low temperatures (Xue 2003). An example of a putatively novel Zn-Deficiency responsive element is the Motif 39 NA (see **Table 3.8**), it is similar to the CRE C1GMAUX28, reported to be involved in the regulation of auxin-related genes (Nagao et al. 1993).

Table 3.8: Examples of predicted motifs putatively responsive to EFTu, Pb oversupply and Zn-Deficiency with the most similar motif from the place database.

Familial profile sequence logo	Motif, Source, name, e-value of similar CRE and stress	Sequence logo of similar CRE
	Motif 80 NZ, Place WBBOXPCWRKY1 8.7705e-10 EF-Tu	
	Motif 55 NA, Place ANAERO2CONSENSUS 3.9082e-05 EF-Tu	
	Motif 64 YA, Place IDE1HVIDS2 1.9026e-08 Pb-Oversupply	
	Motif 349 YZ, Place DRE1COREZMRAB17 1.4526e-07 Pb-Oversupply	
	Motif 44 YA, Place CRTDREHVCBF2 6.5099e-08 Zn-Deficiency	
	Motif 39 NA, Place C1GMAUX28 6.4206e-06 Zn-Deficiency	

Overall, the bioinformatic identification and validation provided by the *in silico* expression analysis yielded a large number of putatively functional CREs. However, the number of distinct elements observed after the STAMP analyses showed that there is a high similarity among the predicted motifs. In order to increase the diversity among the predicted CREs a new approach was followed, the results are described in the next chapter.

3.1.4 Improved *cis*-regulatory elements selection

As an approach to further improve the predictions of novel putatively functional CREs described in **Chapter 3.1.3,** a new set of input sequences was used for the *in silico* expression analysis. The set was comprised of all possible combinations of DNA 10mers, which correspond to 1,048,576 different 10mer input sequences. An *in silico*

expression analysis was run with all possible 10mers and using the results of that analysis, a newly developed tool (see **Chapter 2.6**) identified CRE sets putatively responsive to the stresses stored in the PathoPlant database. To test the approach two sets of CREs putatively responsive to Flg22 and Drought stresses were predicted. They were of special interest as important representatives for a biotic and for an abiotic stress. The *in silico* expression analysis was expanded to include new methods for assessing element specificity and similarity information was also used for element prediction. Using the *cis*-regulatory element finder tool described in **Chapter 2.6,** a biotic set comprised of sequences putatively responsive to Flg22 and *P. syringae* pv. tomato was identified. In a similar way an abiotic set with sequences putatively responsive to Drought and Osmotic stresses was also identified. Although the interest was on Flg22 and Drought responsive sequences, *P. syringae* pv. tomato and osmotic-stress responsive sequences were included in the analysis in order to have a larger set of abiotic and biotic sequences to compare with a similarity analysis (explained later).

Information regarding element specificity, i.e. other possible stresses the predicted sequences can be responsive to, was calculated for each CRE. This was done in order to detect highly specific sequences towards biotic or abiotic stresses. For this purpose sequences from the biotic set showing possible pathway crosstalks with abiotic stresses were filtered out. In the same way sequences from the abiotic set which showed possible pathway crosstalks with biotic-related stresses were also filtered out. The remaining sequences were sorted according to their p-values (from the stress of interest) and the top 30 sequences were chosen for further analysis. Using the STAMP web server a similarity tree with the abiotic and biotic sets was constructed (see **Figure 3.5**). The tree was used to determine which sequences from the biotic set displayed no similarities with abiotic sequences and in the same way which biotic-responsive sequences showed no similarities with abiotic sequences. This was expected to increase the specificity of the predicted sequences towards biotic or abiotic stresses.

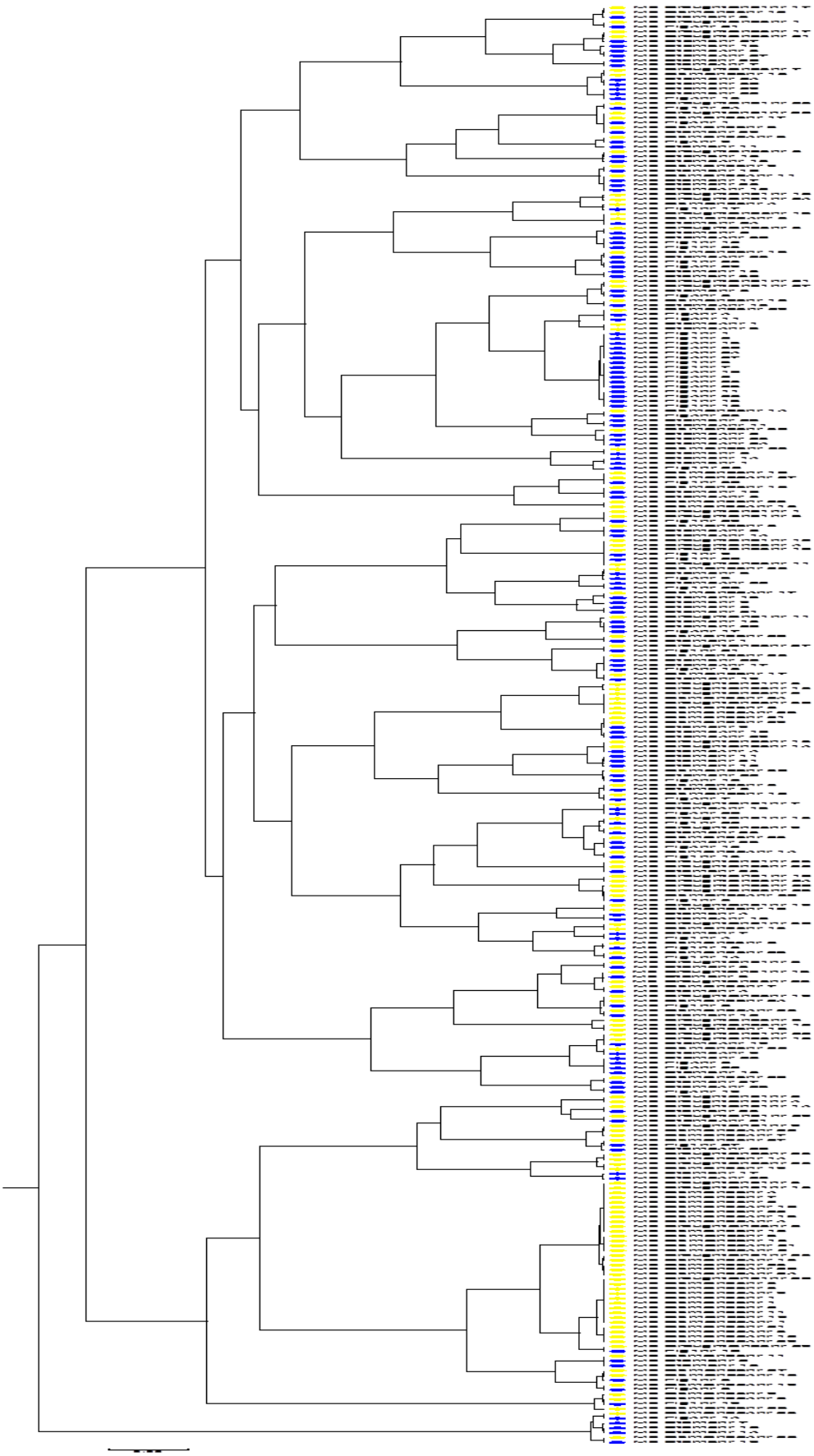


Figure 3.5: Similarity tree between biotic and abiotic sequence sets. Abiotic-responsive and biotic-responsive sequences are marked with a yellow and blue circle, respectively.

The tree in **Figure 3.5** shows some large clusters with only abiotic responsive sequences at the bottom. Also the majority of biotic responsive sequences seem to be at the top of the tree. In total the similarity tree contains 80 clusters, from them, 51 clusters contain a mixture of abiotic and biotic-responsive sequences. In addition 16 clusters were identified containing only biotic-related sequences. These clusters were comprised of 23 Flg22 and 29 P. syringae responsive sequences. Furthermore 13 clusters contained only abiotic-related sequences, which correspond to 22 Drought and 41 Osmotic responsive sequences. Sequences putative responsive to Flg22, which showed no similarities with abiotic responsive sequences, are shown in **Table 3.9**.

Table 3.9: Predicted sequences putatively responsive and highly specific to Flg22 stresses. The time point upon which the sequences from the second column are expected to be responsive to is given in the first column. The number of genes containing the sequence within promoters and the p-value of the expression of such genes under the given stress are shown in the third and fourth column.

Stress	Sequence	Genes	p-value
Flg22 4hr	caaagtcaaa	208	8.39E-15
Flg22 4hr	aaagtcaact	167	1.59E-11
Flg22 4hr	gtcaacgaga	53	1.99E-11
Flg22 1hr	gtcaactcta	36	1.94E-10
Flg22 4hr	gcaaagtcaa	72	3.46E-09
Flg22 4hr	acaaagtcaa	160	5.17E-09
Flg22 4hr	ctcgttgaca	26	5.82E-09
Flg22 1hr	atagttgacc	34	6.04E-09
Flg22 4hr	gtcaacgata	28	1.08E-08
Flg22 4hr	attgtttgac	93	1.97E-08
Flg22 1hr	ggtcaaaaaa	128	1.01E-07
Flg22 1hr	gaggtcaaac	26	1.06E-07
Flg22 1hr	ccaagttgta	32	1.30E-07
Flg22 4hr	gtcaacgtta	37	1.75E-07
Flg22 4hr	aaatctcgat	59	2.50E-07
Flg22 4hr	aactcgaaat	73	3.28E-07
Flg22 1hr	aagtcaacgc	27	1.73E-06
Flg22 1hr	catttttgac	81	1.98E-06
Flg22 4hr	attagagttg	73	2.08E-06
Flg22 1hr	tagatacaca	59	4.49E-06
Flg22 1hr	cggcgagacg	21	6.55E-06
Flg22 1hr	agatcaccaa	50	1.83E-04
Flg22 4hr	gttagagtta	34	2.50E-04

A new similarity tree with the sequences displayed in **Table 3.9** was constructed and the novelty of the predicted sequences was assessed. **Figure 3.6** displays the similarity

tree and familial profiles of each cluster forming the tree. A comparison of all sequences against the databases Agris, AthaMap and Place, point out their similarities with known CREs (see **Table 3.10**). The 14 sequences comprising the largest tree cluster contain the core W-Box motif, which is present in a large number of pathogen-related CREs (Eulgem et al. 2000).

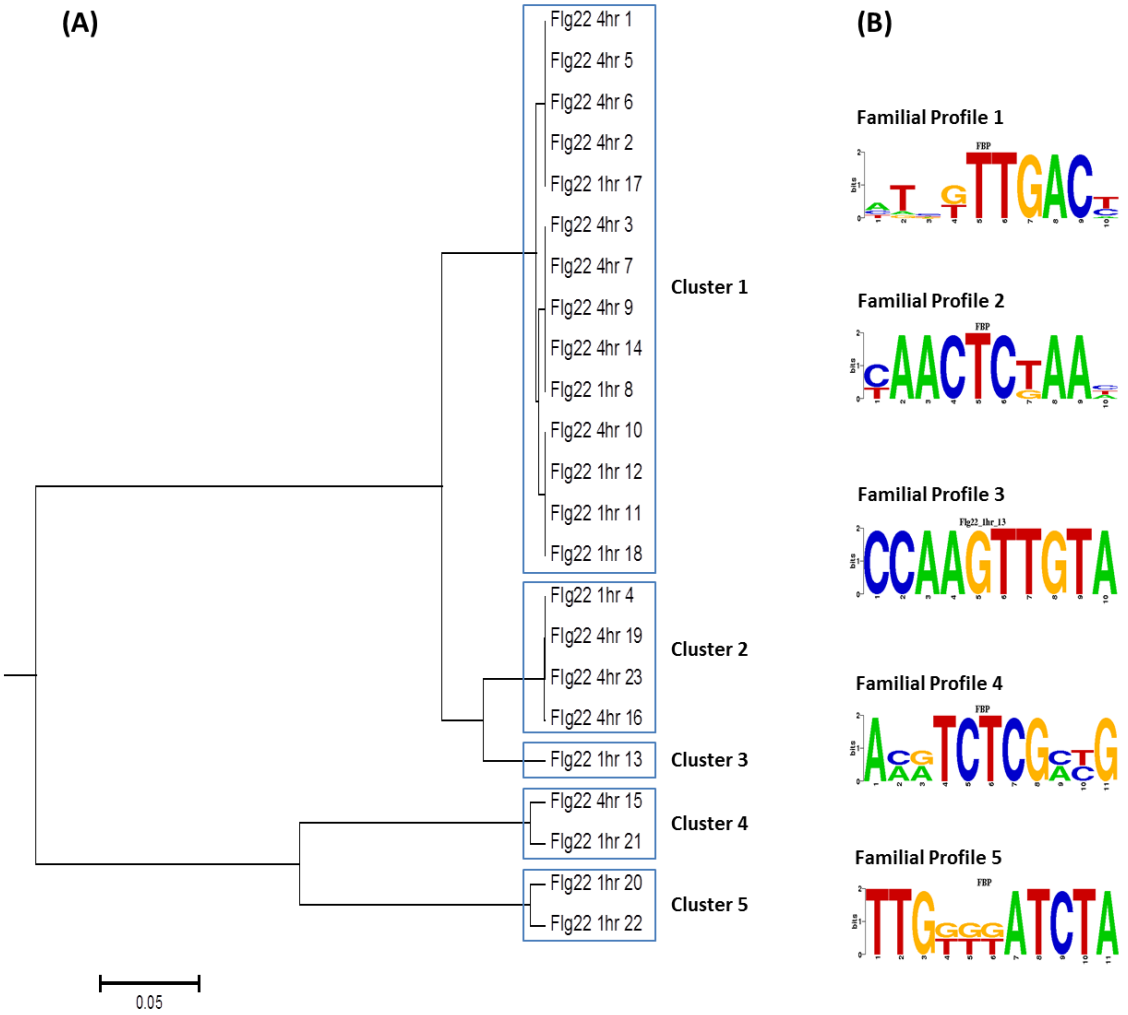
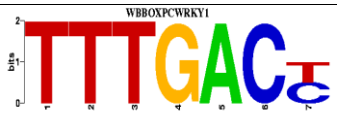
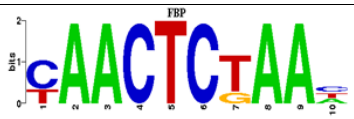
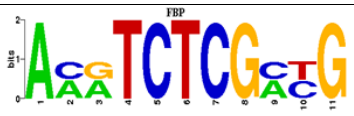
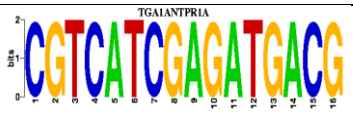
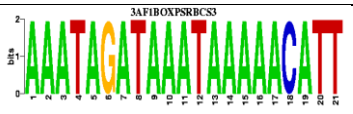


Figure 3.6: Similarity tree (A) and familial binding profiles (B) of sequences putative responsive to Flg22. The clusters were chosen according to their branches length.

Table 3.10: Comparison of familial profiles from the Flg22 set against Agris, AthaMap and Place databases.

Familial profile sequence logo	Cluster, Source, name and e-value of similar CRE	Sequence logo of similar CRE
	Cluster 1, Agris W box 2.9652e-05	
	Cluster 1, AthaMap WRKY6_oneSite 9.2618e-07	
	Cluster 1, Place WBBOXPCWRKY1 1.4621e-06	
	Cluster 2, Place CAREOSREP1 3.4151e-07	
	Cluster 2, Agris TELO-box, 1.5778e-05	
	Cluster 3, Place RBENTGA3 4.5575e-12	
	Cluster 4, Place TGA1ANTPR1A 1.0270e-07	
	Cluster 5, Place 3AF1BOXPSRBCS3 3.1785e-07	

As observed in **Table 3.10** sequences comprising cluster 2 display similarities with CAREOSREP1, a CRE from the database Place which has been suggested to be the regulator of hydrolase gene expression induced by gibberellins (Sutoh and Yamauchi 2003). The CREs from cluster 2 also show similarities to the TELO-box CRE which has been shown to be involved in the control of gene expression and in the activation of the Elongation Factor 1 Alpha (eEF1A) (Tremousaygue et al. 1999). A very low e-value, i.e. a very high similarity to the CRE RBENTGA3 was observed for the single sequence forming cluster 3. This CRE has been shown to serve as the binding site of a transcriptional activator that plays a role in regulation of cell elongation by controlling the quantity of the hormone gibberellin (Fukazawa et al. 2000). The sequences comprising cluster 4 have a high similarity to the CRE TGA1ANTPR1A, which was shown to enhance the expression of a pathogen-related gene (Strompen et al. 1998). Finally

cluster 5 is formed by sequences displaying similarities to 3AF1BOXPSRBCS3, a CRE involved in light responses (Lam et al. 1990).

Table 3.11: Predicted sequences putative responsive and highly specific to Drought stresses. The exact time point upon which the sequences from the second column are expected to be responsive to is given in the first column. The number of genes containing the sequence within promoters and the p-value of the expression of such genes under the given stress are shown in third and fourth column.

Stress	Sequence	Genes	p-value
Drought 1hr	agtgagtggg	22	2.33E-09
Drought 1hr	acacgtggcc	36	3.68E-09
Drought 3hr	gaaacgattc	54	4.48E-09
Drought 1hr	cacgtggcca	24	1.10E-08
Drought 3hr	ccttcgttta	28	2.97E-08
Drought 1hr	acccactcac	22	3.90E-07
Drought 1hr	ataacggtta	33	4.54E-07
Drought 3hr	aagcacgtgg	22	9.45E-07
Drought 3hr	aaaggagcct	20	1.30E-06
Drought 1hr	aaagtaccct	34	1.54E-06
Drought 1hr	aactagctag	65	3.79E-06
Drought 3hr	ctaagcctga	21	5.98E-06
Drought 3hr	gagcctctaa	22	1.23E-05
Drought 1hr	cgtgtcactc	27	1.77E-05
Drought 1hr	ggagaataca	25	1.78E-05
Drought 1hr	cacactccaa	30	1.84E-05
Drought 3hr	aaaggttcgt	22	2.01E-05
Drought 1hr	gcgacgtttc	30	5.05E-05
Drought 3hr	tcacttcgga	27	1.05E-04
Drought 1hr	caagagtata	67	1.23E-04
Drought 1hr	caaccatccc	20	3.91E-04
Drought 3hr	acacttgtga	52	4.19E-04

As explained before, the tree in **Figure 3.5** contained 13 clusters comprised only of abiotic-responsive sequences, which corresponded to 22 Drought (see **Table 3.11**) and 41 Osmotic responsive sequences. The Drought set was used to construct a similarity tree, where the sequences were grouped into clusters which in turn were used to generate familial profiles (see **Figure 3.7**). Such profiles were compared with known CREs stored in the databases Agris, AthaMap and Place. The most similar CREs are presented in **Table 3.12**.

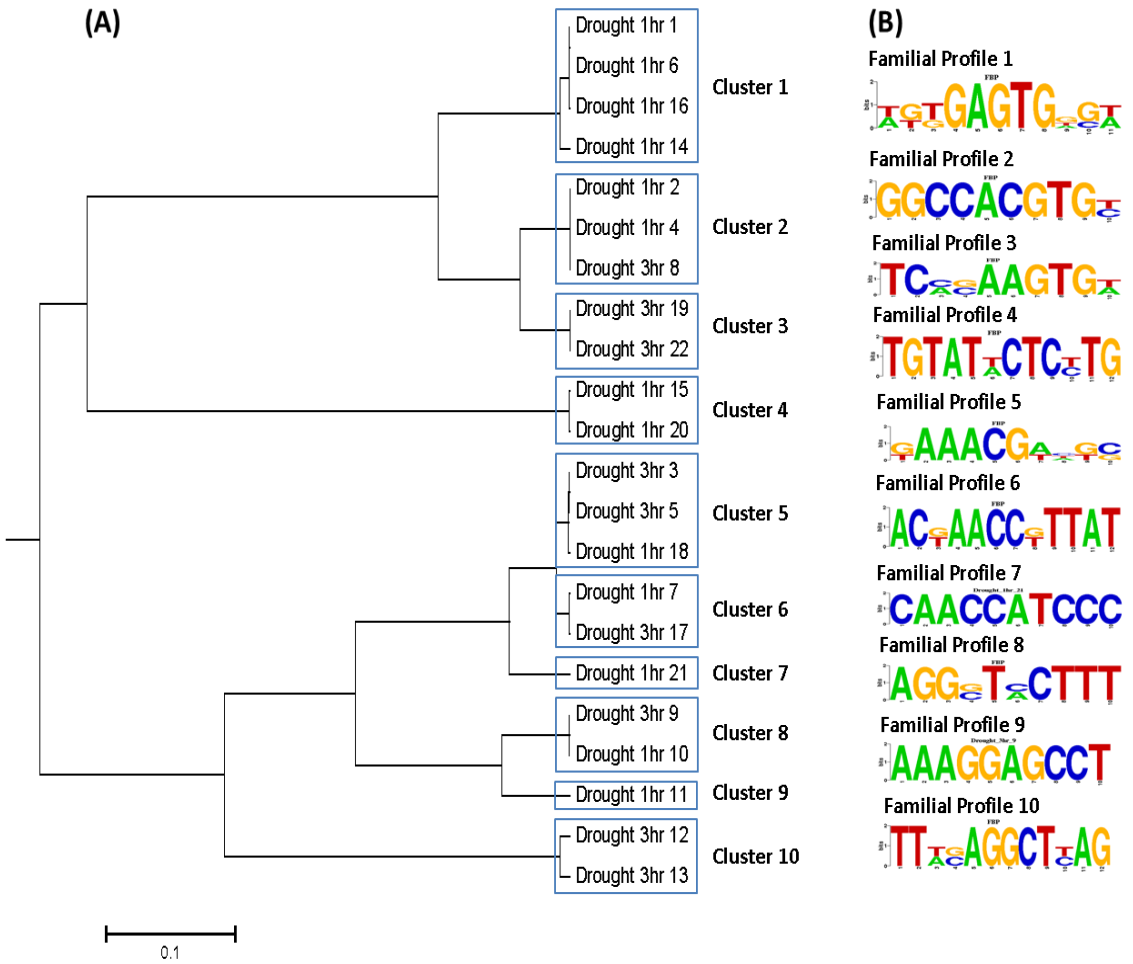
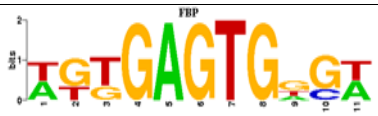
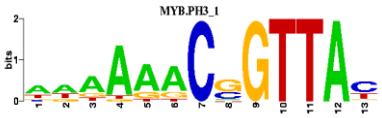
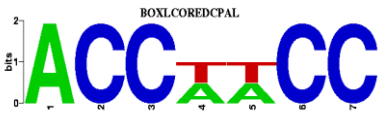


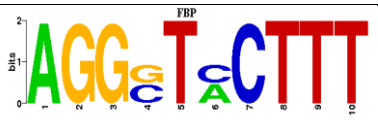
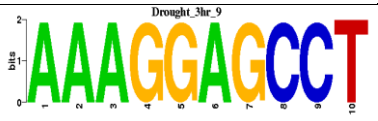
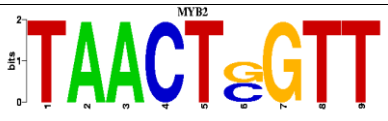
Figure 3.7: Similarity tree (A) and familial binding profiles (B) of sequences putative responsive to Drought. The clusters were chosen according to their branches length.

Sequences grouped into cluster 1 were found to have a high similarity to the CRE SORLIP5AT from the Place database. It is a CRE reported to be overrepresented in the promoters of light-induced genes (Hudson and Quail 2003). Sequences comprising cluster 2 display very high similarities to abscisic acid-related CREs from the AGRIS and AthaMap databases. The CRE SORLIP3AT, identified as playing a role in the gene expression regulation of phytochrome-A (Hudson and Quail 2003), was observed to be similar to the sequences from cluster 3. Sequences forming cluster 4 display similarities with a CRE named OCSGMHSP26A that is very important for the activity of the ocs-element, which regulates the expression pathogen-related genes (Ellis et al. 1993). ABRE3HVA1, another abscisic acid-responsive element (Shen et al. 1996) was found to be similar to sequences from cluster 5. Sequences comprising cluster 6 and 9 seem to have similarities with MYB-related CREs. The elements include, AtMYB2 from Agris, a CRE that has been shown to be involved in responses to dehydration (Abe et al. 1997),

MYB.PH3_1 from AthaMap, which plays a role in the flavonoid biosynthesis (Solano et al. 1995) and MYB1LEPR from place, involved in the regulation of gene-expression in defense responses (Chakravarthy et al. 2003). The single sequence of cluster 7 is similar a CRE (BOXLCOREDCPAL) involved in gene expression as a response to environmental stresses (Maeda et al. 2005). Sequences comprising cluster 8 are similar to the wound responsive element (Palm et al. 1990) WRECSAA01. Finally no significant similarities were observed for sequences forming cluster 10.

Table 3.12: Comparison of familial profiles from the Drought set against similar CREs from the Place, Agris and AthaMap database.

Familial Profile	Source, name and e-value of similar CRE	Sequence logo of similar CRE
	Cluster 1, Place SORLIP5AT 8.0933e-06	
	Cluster 2, Agris ABFs 1.0971e-11	
	Cluster 2, AthaMap ABF1 5.2828e-12	
	Cluster 3, Place SORLIP3AT 1.9433e-06	
	Cluster 4, Place OCSGMHSP26A 5.3043e-05	
	Cluster 5, Place ABRE3HVA1 2.2185e-06	
	Cluster 6, Agris AtMYB2 6.5743e-05	
	Cluster 6, AthaMap MYB.PH3_1 5.0483e-06	
	Cluster 6, Place MYB1LEPR 6.5743e-05	
	Cluster 7, Place BOXLCOREDCPAL 3.6117e-07	

	Cluster 8, Place WRECSAA01 9.2021e-05	
	Cluster 9, Agris MYB2 4.0346e-05	
	Cluster 9, AthaMap MYB.PH3_2 8.5777e-05	
	Cluster 9, Place MYB1LEPR 1.2359e-05	
	Cluster 10 No significant similarity	N.A.

Overall several CREs putatively responsive to Flg22 and Drought were predicted. The fact that sequences similar to known CREs involved in pathogen and dehydration responses were found among the predicted Flg22 and Drought sequence sets suggests that the sequences are putatively functional. Furthermore new putatively functional CREs were also present in the predicted sets, suggesting that the method serves to predict a high variety of CREs.

3.1.5 Novel web-tools for *cis*-regulatory element prediction

Two web-tools were developed in the course of the present study, an on-line version of the *in silico* expression analysis and another tool for *cis* elements prediction. The *in silico* expression analysis is freely available at http://www.pathoplant.de/expression_analysis.php. The web tool allows the validation of a single sequence as a putatively functional CRE responsive to biotic and abiotic stresses. The analysis starts by providing the DNA sequence expected to be a CRE in the text box under the label “Potential *cis*-sequence”. The sequence will be searched in the *Arabidopsis thaliana* gene promoters, whose length can be selected by the user (either 500 or 1000 nucleotides). In the course of the present study PathoPlant updated their TAIR release from 7 to 8, therefore TAIR release 8 sequence and annotation data are used by the on line tool as well. Genes putatively regulated by smallRNAs and miRNAs can be excluded from the analysis by activating the

corresponding checkboxes (see **Figure 3.8**). The button “Demo” enters an example sequence into the potential *cis*-sequence text box, and the “Reset” button clears any text entered in the text box. The search is started by clicking the “Search” button.

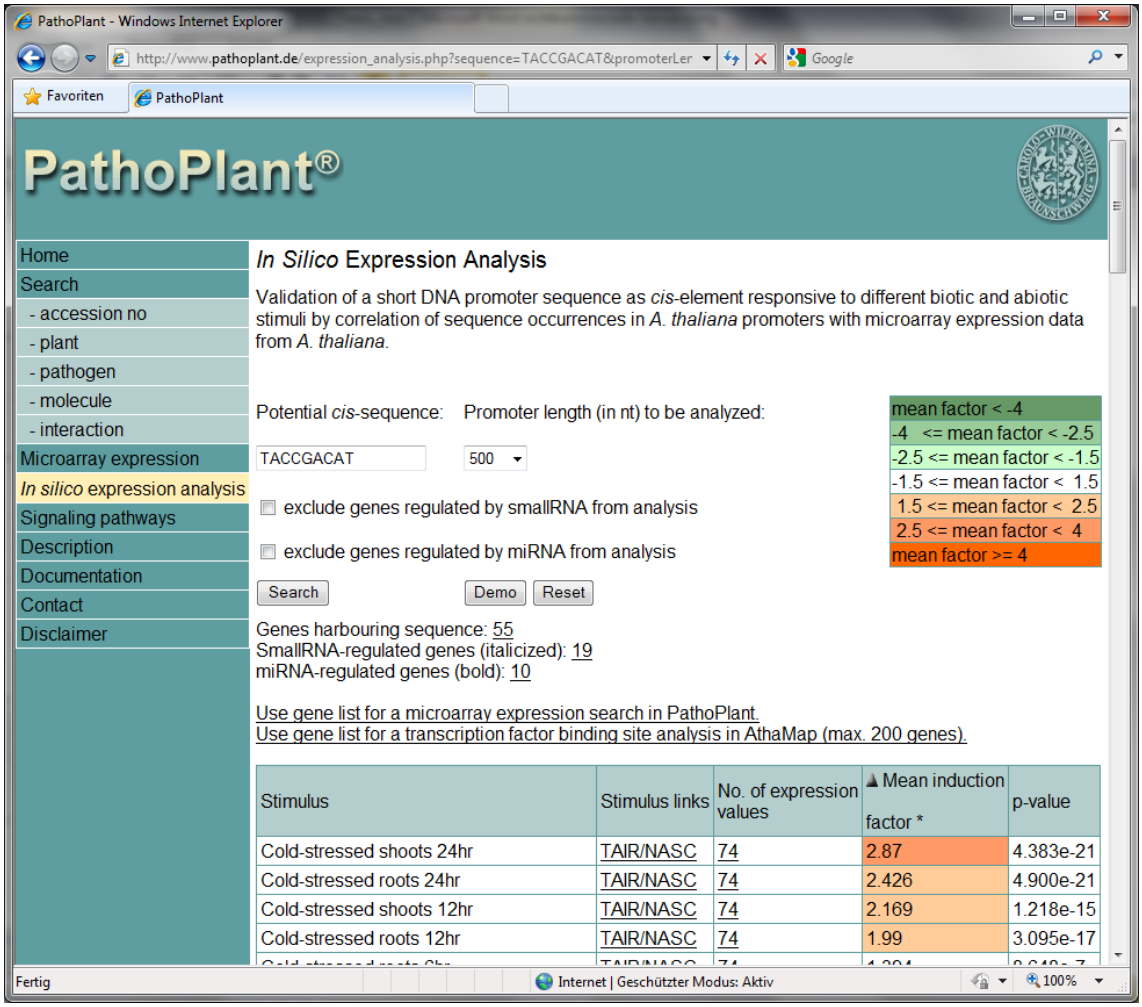


Figure 3.8: Screenshot of the in silico expression anaylsis web tool. The results obtained after starting a search with the abiotic stress-responsive sequence TACCGACAT are shown. The promoter length can be selected for each search. A description (not shown) of the genes containing the sequence within promoters can be displayed by selecting the number of ‘Genes harboring sequence’. In addition the number of genes putatively regulated post-transcriptionally by smallRNAs and miRNAs can also be displayed. A table is shown containing the microarray experiment (stimulus) with corresponding source link, the number of expression values available for genes harboring the sequence, the mean induction factor of those genes, color-coded to visually identify the strength of the up- or down-regulation and a p-value indicating the significance of that value. Values can be sorted according to stress, mean induction factor and p-value.

After performing a search, the number of genes containing the entered sequence within promoters is shown as a result. This number can be clicked to display a column with a list of genes harboring the sequence within the promoters. An additional click to

the arrow on the right side of the column header displays the description of the genes containing the sequence in the promoters (see **Figure 3.9**). Similar lists are shown by selecting the number of smallRNA- and miRNA-regulated genes. Found genes can be used to perform a microarray expression search in PathoPlant, where the individual expression of each gene under all stresses stored in the database is shown. Additionally the gene list can also be used to perform a transcription factor binding site analysis in AthaMap, which will show known transcription factor binding sites within the promoters. **Figure 3.8** also shows a table containing: the stress name (*Stimulus*), a link to the microarray experiment source (*Stimulus links*), the number of expression values in the experiment used to calculate the average expression (*No. of expression values*), the average expression (*Mean induction factor*) and the statistical significance of the average expression (*p-value*). The table is sorted by default according to the column *Mean induction factor* and can be resorted according to the *Stimulus, Mean induction factor* and *p-value* columns.

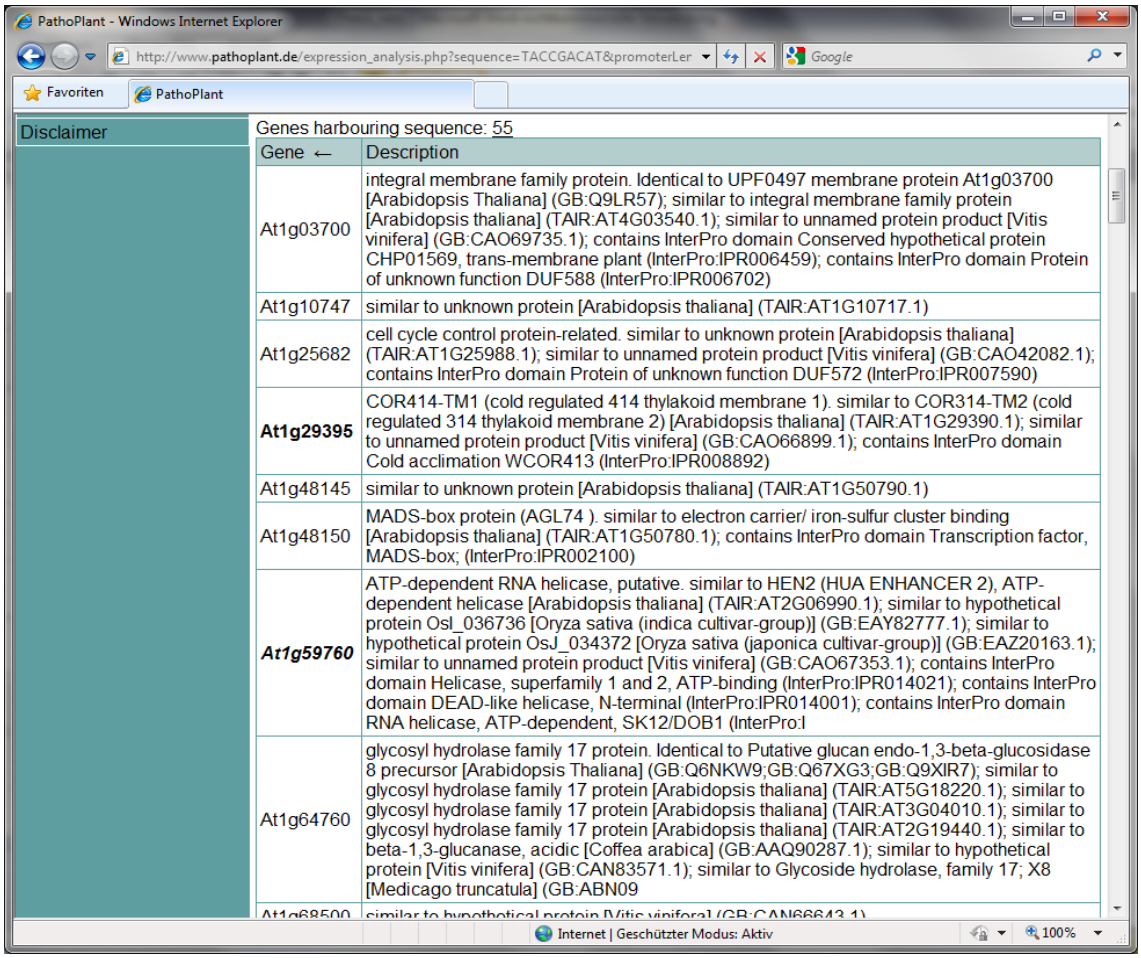


Figure 3.9: Screenshot of gene details shown in the *in silico* expression analysis on-line tool after a search.

Details about the positions within the promoters where the potential CRE occurs, as well as individual expression values for a given stress are obtained by clicking the numbers in the column *No. of expression values* (see **Figure 3.8** and **Figure 3.10**). The details are shown in a new window or tab which displays a table with: gene names, sequence orientation within promoters, sequence relative distance to the point of reference (TSS if known, otherwise the ATG), induction factor values for a gene upon the selected stress, mean of the induction factors and the number of replicates (n) and the base-10 logarithm of the standard deviation for mean induction factor. The on-line tool currently uses TAIR release 8 sequence and annotation data.

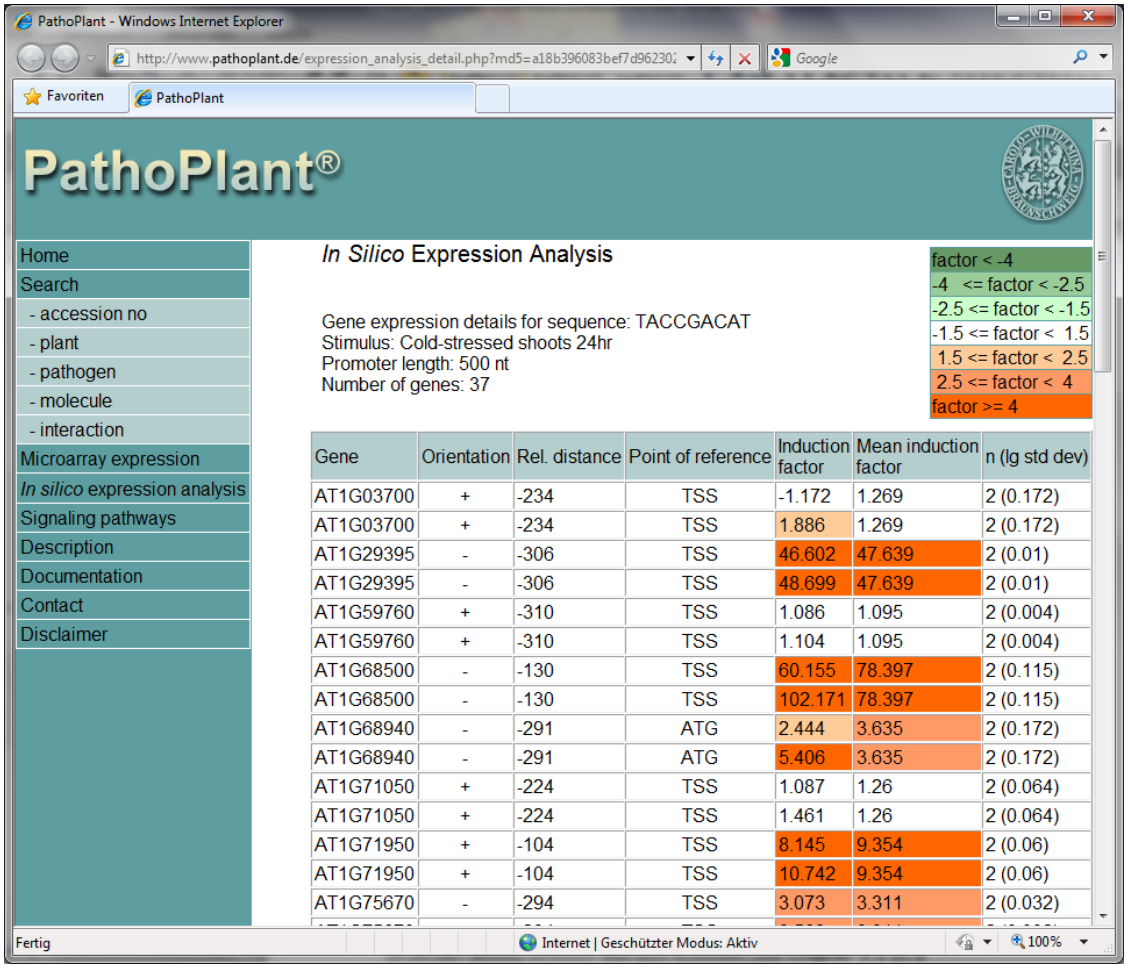


Figure 3.10: Screenshot showing sequence positional information within promoters and individual gene expression values for a given stress after an *in silico* expression analysis is performed.

Another tool developed in the present study was *cis*-elements, which will be integrated into PathoPlant following a publication at http://www.pathoplant.de/. The tool allows the identification of putatively functional DNA 8, 9 and 10mers responsive to biotic and

abiotic stresses. *cis*-elements uses the output of an *in silico* expression analysis performed with all possible DNA 8, 9 and 10mers (using TAIR release 8 sequence and annotation) in order to identify putatively functional CREs responsive to a selected stress. The identified CREs meet the selection criteria described in **Chapter 2.6**, i.e. a minimum of 20 genes should contain the sequence within promoters, genes should be up-regulated (i.e. the must have a mean induction factor above 1.0) upon selected stress and the p-value should be $\leq$ 0.001. In addition the tool allows defining specificity parameters for CRE selection.

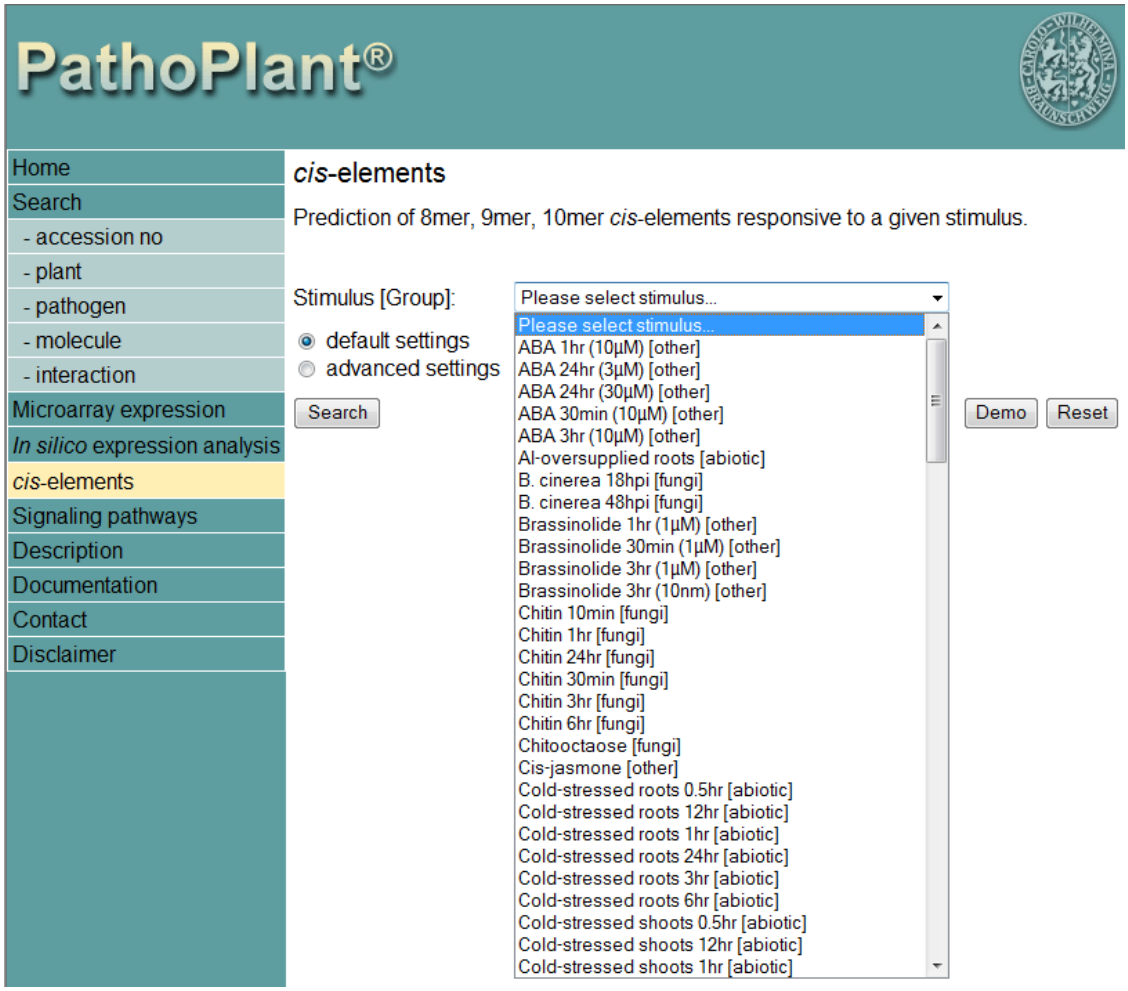


Figure 3.11: Screenshot of the web tool *cis*-elements. A stress can be selected in order to retrieve putatively functional CREs. Specificity of predicted elements can be defined by using default or advanced settings.

A screenshot of the tool (off line) is shown in **Figure 3.11**. By clicking the label "Please select stimulus…" it is possible to select a stress to predict putatively functional CREs. The name of each stress is followed by a group name enclosed in brackets. The group corresponds to the type of the selected stress, which include: abiotic, biotic (excluding

fungi), fungi and other (including phytohormones and developmental stresses). The specificity of the predicted CREs can be defined with either default or advanced settings. With default settings, CREs putatively responsive to the selected stress and to other stresses of the same type will be predicted. E.g. by selecting the stress Cold-stressed roots 24hr, predicted CREs will be putatively responsive to the selected Cold stress and also to stresses of the group abiotic. The specificity can be further defined by clicking the radio button “advanced settings”, which will show further selection options (see **Figure 3.12**). It is possible to further select the stress types, for which the predicted CREs are putatively responsive to. Selecting any of the check boxes for the search will result in predicted CREs putatively responsive to at least one stress of the selected stress type. Thus, if the stress Cold-stressed roots 24hr is selected and the check boxes “abiotic stimuli”, “fungi stimuli” and “biotic stimuli” are checked, predicted CREs will be responsive to Cold-stressed roots 24hr and also to at least one stress of the types abiotic, fungi and biotic. Alternatively by unchecking “abiotic stimuli”, “fungi stimuli”, “biotic stimuli” and “other stimuli” it is possible to check “exclusive for selected stimulus” which in the example will mean that the predicted CREs are only expected to be responsive to Cold-stressed roots 24hr and no other stress. It is also possible to select the sequence length of the predicted CREs by clicking the drop down menu next to the label “Sequence length (nt)”. A click to the “Search” button starts the CRE prediction.

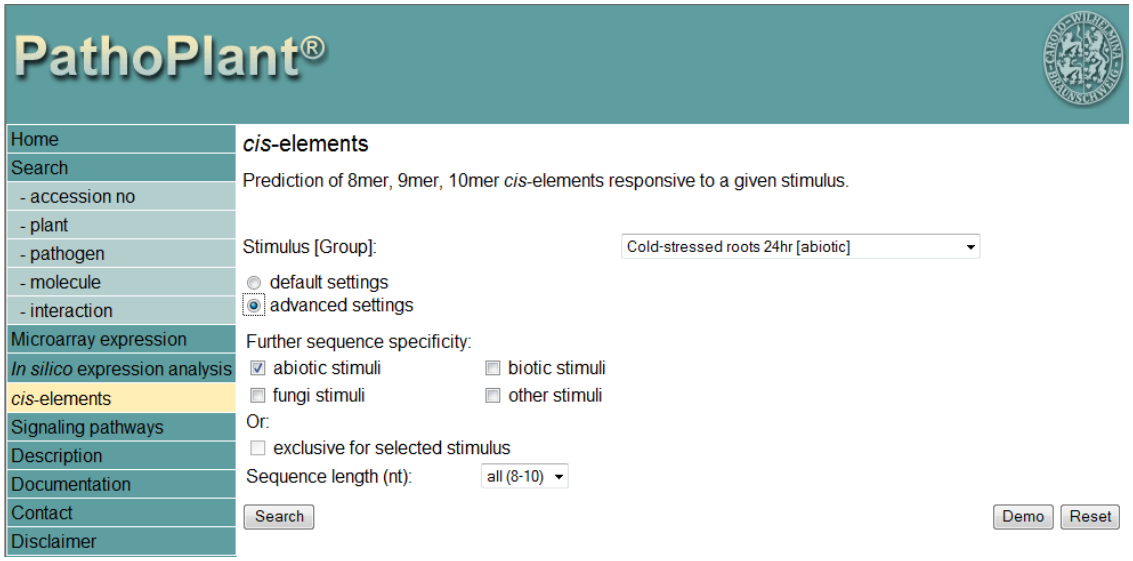
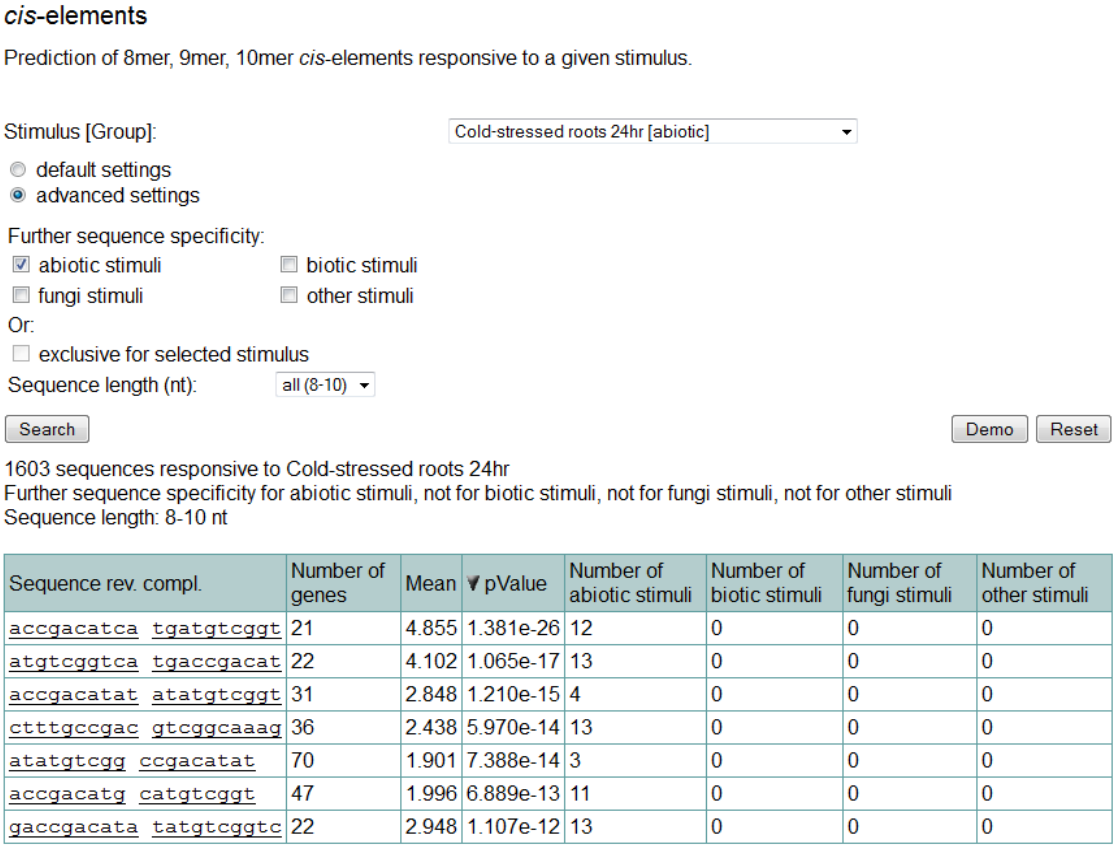


Figure 3.12: Screenshot of advanced specificity settings in the tool *cis*-elements.

The number of predicted CREs, together with the selected sequence specificity and sequence length are shown after clicking the "Search" button (see **Figure 3.13**). The predicted CREs are shown in a table containing: the CREs sequence with its reverse complement, the number of genes containing the sequence within promoters, the mean expression value of such genes upon selected stress, the statistical significance of that mean value and the number of abiotic, biotic, fungi and/or other stresses which also show significant statistical values for that sequence. Thus, the values in the last four columns show the predicted sequence specificity. The results table can be resorted according to each column. Finally, by clicking the CRE sequence or its reverse complementary, an *in silico* expression analysis is performed.



cis-elements

Prediction of 8mer, 9mer, 10mer *cis*-elements responsive to a given stimulus.

Stimulus [Group]: Cold-stressed roots 24hr [abiotic]

default settings
advanced settings

Further sequence specificity:
abiotic stimuli biotic stimuli
fungi stimuli other stimuli
Or:
exclusive for selected stimulus
Sequence length (nt): all (8-10)

Search Demo Reset

1603 sequences responsive to Cold-stressed roots 24hr
Further sequence specificity for abiotic stimuli, not for biotic stimuli, not for fungi stimuli, not for other stimuli
Sequence length: 8-10 nt

Sequence rev. compl.	Number of genes	Mean	pValue	Number of abiotic stimuli	Number of biotic stimuli	Number of fungi stimuli	Number of other stimuli
accgacatca tgatgtcggt	21	4.855	1.381e-26	12	0	0	0
atgtcggtca tgaccgacat	22	4.102	1.065e-17	13	0	0	0
accgacatat atatgtcggt	31	2.848	1.210e-15	4	0	0	0
ctttgccgac gtcggcaaag	36	2.438	5.970e-14	13	0	0	0
atatgtcgg ccgacatat	70	1.901	7.388e-14	3	0	0	0
accgacatg catgtcggt	47	1.996	6.889e-13	11	0	0	0
gaccgacata tatgtcggtc	22	2.948	1.107e-12	13	0	0	0

Figure 3.13: Screenshot of *cis*-elements results.

3.2 Pathway crosstalks

A single transcription factor can be associated to different stresses resulting in signaling pathway crosstalks. Thus, CREs serving as binding sites for such transcription factors are also said to be responsive to such stresses. In order to identify these possible pathway crosstalks, analyses described in **Chapter 3.2.1** were performed.

They determined if predicted CREs responsive to single stresses can also be associated to other stresses. In addition, the degree of overlapping in CREs putatively responsive to certain abiotic stresses was also assessed (see **Chapter 3.2.2**).

3.2.1 Specificity of predicted motifs

Novel putatively functional CREs presented in **Chapter 3.1.3** were predicted to be responsive to single stresses. The specificity of such elements was assessed by performing pathway crosstalk analysis, where it was determined if they can be responsive to further stresses. Pathway crosstalk analyses yielded overall expression values of genes containing predicted CREs within promoters. This information was used to assess other possible stresses (besides the stress, the elements were initially predicted to be responsive to) for which the CREs can also be responsive to. Overall expression values were ranked in tables where stresses showing high overall mean values imply that CREs could also be responsive to these stresses. Furthermore a p-value indicates if the overall expression value for certain stress is very similar to the one of the initial stress.

Table 3.13 shows crosstalk analyses for CREs putatively responsive to Flg22 1hr. The ranking indicates that among the top 10 highest overall mean values, the stresses are mainly biotic and fungal-related. Three abiotic related stresses (salt and osmotic stresses) are also present among the top 10 stresses with highest overall mean values. However, the majority of biotic and fungal stresses indicate that CREs are expected to be mainly responsive towards such type of stresses. In order to identify the most similar stress to Flg 22 1hr the p-values in the table should be assessed. In this case, a high p-value indicates a high similarity to the stress Flg22 1hr. In that way the stresses Harpin Z 1hr and EF-Tu 60min can be identified as the most similar to Flg 22 1hr. Interestingly, the stresses are of the same type (bacterial elicitors) and even have the same time point 1hr. The abiotic stress Salt-stressed roots 6hr also displays some degree of similarity to Flg22 1hr, although not as similar as the above mentioned stresses. These results indicate that predicted CREs for the stress Flg22 1hr also display a putative very similar response to the related stresses Harpin Z 1hr and EF-Tu 60min and a possible abiotic-biotic stress pathway crosstalk.

Table 3.13: Top 10 stresses showing high overall mean values in crosstalk analysis for all predicted CREs putatively responsive to Flg22 1hr. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Flg22 1hr.

Stress	Mean	p-value
Flg22 (P. syringae) 1hr	2.5204	N.A.
Harpin Z 1hr	2.3865	2.188E-01
EF-Tu 60min	2.2099	1.038E-01
Salt-stressed roots 6hr	1.9163	9.104E-04
P. infestans 6hpi	1.7747	1.705E-07
Harpin Z 4hr	1.6855	9.764E-08
EF-Tu 30min	1.6523	7.599E-06
P. syringae pv. phaseolicola 24hpi	1.6208	6.226E-10
Osmotic-stressed shoots 3hr	1.6131	5.046E-08
Salt-stressed roots 24hr	1.5930	7.974E-07
…	…	…

The next time point analyzed for the stress Flg22 was 4hr (see **Table 3.14**). The top 10 stresses showing high overall mean values are only fungal and biotic-related. There is a majority of stresses related to the bacterial pathogen P. syringae. From the ranking list it can also be determined that the stress Harpin Z 4h is the most similar to Flg 4h, although not with a very high p-value. Nevertheless both stresses (Flg22 1h and 4h) show the same stress as the most similar (Harpin Z), which implies that pathway crosstalks can occur, since both stresses could be regulated with similar transcription factors and therefore a similar set of CREs.

Table 3.14: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Flg22 4hr. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Flg22 4hr.

Stress	Mean	p-value
Flg22 (P. syringae) 4hr	2.3489	N.A.
Harpin Z 4hr	2.1820	6.164E-04
P. syringae pv. phaseolicola 6hpi	1.8216	1.977E-20
P. syringae pv. phaseolicola 24hpi	1.7759	2.248E-18
P. syringae pv. tomato hrcC- 24hpi	1.7433	4.096E-20
P. syringae pv. tomato hrcC- 6hpi	1.5736	3.825E-43
Flg22 (P. syringae) 1hr	1.5513	4.826E-43
Harpin Z 1hr	1.5012	1.340E-52
NPP1 (P. parasitica) 4hr	1.4576	3.199E-51
EF-Tu 60min	1.3938	2.371E-23
…	…	…

Pathway crosstalks between biotic and abiotic stresses were also observed for CREs putatively responsive to Chitooctaose (see **Table 3.15**). Among the top 10 highest overall means, a majority of abiotic stresses (mainly Salt-stressed roots at different time points) are present, implying that predicted CREs display also a putative responsiveness to abiotic stresses. This is further observed by assessing the p-values shown in the table where stresses EF-Tu 60min and Salt-stressed roots 6hr are identified as the most similar to Chitooctaose. Similarities between EF-Tu and Salt stresses were also observed in **Table 3.13** and when analyzing CREs putatively responsive to EF-Tu (see **Table 7.3** and **Table 7.4** in **Chapter 7.7**). Together the results indicate a possible pathway crosstalk between Chitooctaose, EF-Tu and Salt stressed roots stresses.

Table 3.15: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Chitooctaose. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Chitooctaose.

Stress	**Mean**	**p-value**
Chitooctaose	2.5555	N.A.
EF-Tu 60min	2.3316	1.483E-01
Salt-stressed roots 6hr	2.1710	5.099E-02
EF-Tu 30min	1.8840	5.304E-05
Salt-stressed roots 3hr	1.7728	8.306E-06
Osmotic-stressed shoots 1hr	1.6700	4.308E-07
Salt-stressed roots 12hr	1.6256	1.658E-08
Cold-stressed shoots 3hr	1.5839	9.022E-06
Salt-stressed roots 24hr	1.5796	5.070E-08
Osmotic-stressed shoots 3hr	1.5310	4.915E-09
…	**…**	**…**

CREs predicted for analyzed abiotic stresses showed a very specific responsiveness. Overall mean expression values calculated for CREs putatively responsive to Pb-oversupplied 25ppm leaves are show on **Table 3.16**. The majority of overall mean values for other stresses is low, only the same stress with a different concentration (Pb-oversupplied 50ppm leaves) displays a high overall mean value which suggests that there are no pathway crosstalks with other stresses. In addition the extremely low p-values also indicate that the CREs should be very specific to the stress Pb-oversupplied 25ppm leaves. All the Pb-related stresses display similar values for crosstalk analysis (see **Chapter 7.7**).

Table 3.16: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Pb 25ppm leaves. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Pb 25ppm leaves.

Stress	Mean	p-value
Pb-oversupplied (25ppm) leaves	2.8839	N.A.
Pb-oversupplied (50ppm) leaves	1.9929	2.094E-21
Pb-oversupplied (50ppm) roots	1.2317	3.778E-105
Pb-oversupplied (25ppm) roots	1.1654	1.313E-136
Inflorescence vs. young leaves	1.0484	1.122E-271
Inflorescence vs. shoot apex, vegetative	1.0373	1.073E-277
Zn-deficient roots	1.0289	5.194E-295
Osmotic-stressed roots 0.5hr	1.0244	3.214E-293
Methyl-jasmonate 1hr	1.0189	7.645E-288
Cold-stressed shoots 0.5hr	1.0164	6.483E-296
…	…	…

A similar stress-specific response was observed for CREs putative responsive to Zn stresses (see **Table 3.17** for deficiency, see **Chapter 7.7** for oversupply and resupplied Zn) where mainly Zn-related stresses where identified with significant similarities. Thus, the results suggest that, for the analyzed stresses, the abiotic CREs seem to be more specific that the biotic ones.

Table 3.17: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-deficient roots. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-deficient roots.

Stress	Mean	p-value
Zn-deficient roots	2.6532	N.A.
Zn-resupplied roots 2hr vs. suffficient Zn	1.9015	7.065E-08
Zn-deficient shoots	1.5060	6.992E-11
Salt-stressed roots 24hr	1.3032	3.762E-24
Zn-resupplied shoots 8hr vs. suffficient Zn	1.3010	1.281E-20
Chitin 3hr	1.2992	4.374E-17
Salt-stressed roots 6hr	1.2841	6.279E-20
Chitin 6hr	1.2547	9.203E-20
Pb-oversupplied (25ppm) roots	1.2010	8.319E-19
Pb-oversupplied (50ppm) roots	1.1614	1.038E-17
…	…	…

3.2.2 Specificity of abiotic stress responsive motifs

As explained in the last chapter, biotic-abiotic pathway crosstalks were observed. Nearly always abiotic stresses salt, osmotic and cold were present in the ranked lists. In order to have a closer look at this abiotic response, all possible CREs responsive to these abiotic stresses were identified and the similarities among these elements were assessed (see the electronic appendix **E_3.2.2** for a complete list of predicted sequences for each stress). It has also been reported that CREs responsive to these abiotic stresses (salt, osmotic and cold) are also responsive to Drought stress (Yamaguchi-Shinozaki and Shinozaki 1994), therefore CREs putatively responsive to Drought stress were also included in the analysis (see the electronic appendix **E_3.2.2** for the corresponding sequences). All possible CREs responsive to the abiotic stresses were identified using the tool described in **Chapter 2.6**), which predicted CRE sets putatively responsive to the abiotic stresses Osmotic, Cold, Salt and Drought. In order to visualize the sequence similarities and the number of CREs among the predicted sets, area- proportional venn diagrams were generated. Such Venn diagrams allowed a graphical representation of overlapping CREs among the sets, which in turn will allow the identification of convergence points in signaling pathways of the analyzed stresses. The diagrams were generated for each available stress time point in PathoPlant (0.5hr, 1hr, 3hr, 6hr, 12hr and 24hr), i.e. Cold, Osmotic, Salt and Drought stresses where compared with each other at each time point. **Figure 3.14** displays such sequence comparison for time point 0.5hr. The diagram shows that CREs responsive to Drought-stressed shoots are the majority among the analyzed elements. In addition, overlapping elements are present among all sets, which suggest a common early regulation of these abiotic stresses.

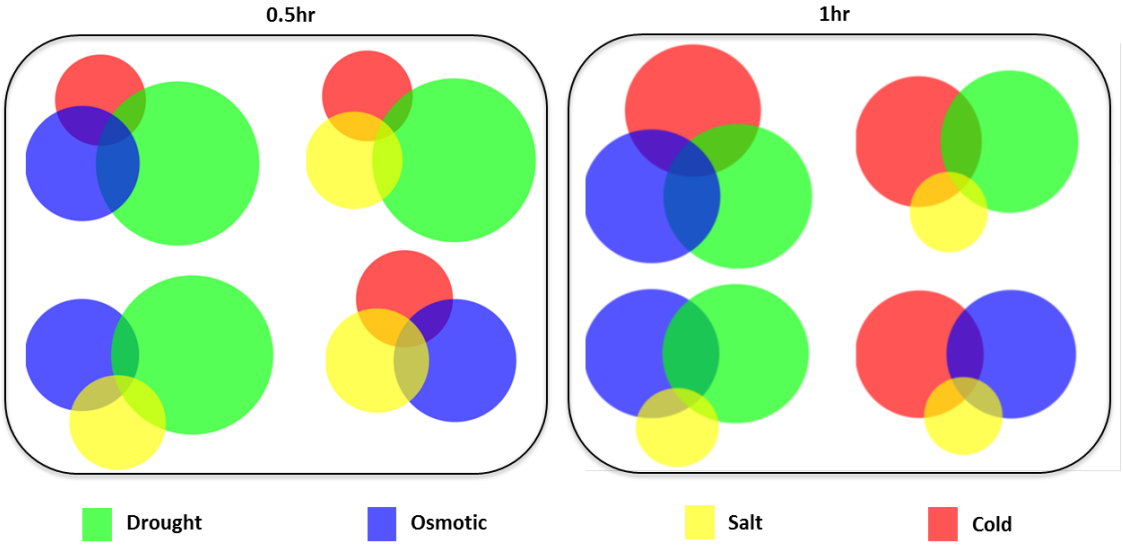


Figure 3.14: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 0.5hr and 1hr Shoots putatively responsive sets. Each circle represents a CRE set.

The element proportions changes at time point 1hr as can be seen on the right side of **Figure 3.14**. Although the Drought CREs remain a majority among the sets, the differences are no longer as high as observed for time point 0.5hr, which means that at this time point the number of Osmotic and Cold-responsive elements increases. The sets still show a high overlapping number of CREs, which, as shown earlier, is also present for CREs at time point 0.5 hr.

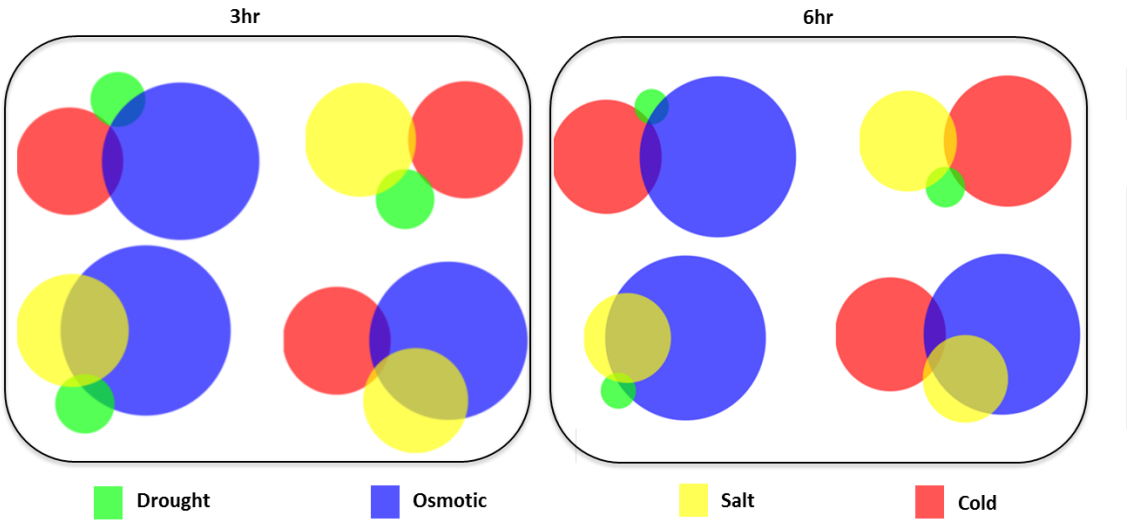


Figure 3.15: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 3hr and 6hr Shoots putative responsive sets. Each circle represents a CRE set.

The amount of putative Drought responsive CREs is dramatically reduced at time points 3hr and 6hr in comparison with the other stresses (see **Figure 3.15**). The proportions change and the number of putative Osmotic stress responsive CREs become the majority. A high overlapping of sequences can be observed between Osmotic and Salt stresses and, on the contrary, almost no overlapping between Cold, Drought and Salt stresses. A similar proportion of CREs is observed at time points 12hr and 24hr (see **Figure 3.16**). The putative Osmotic responsive CREs are still the majority and the Drought CREs are even further reduced in comparison with the other stresses.

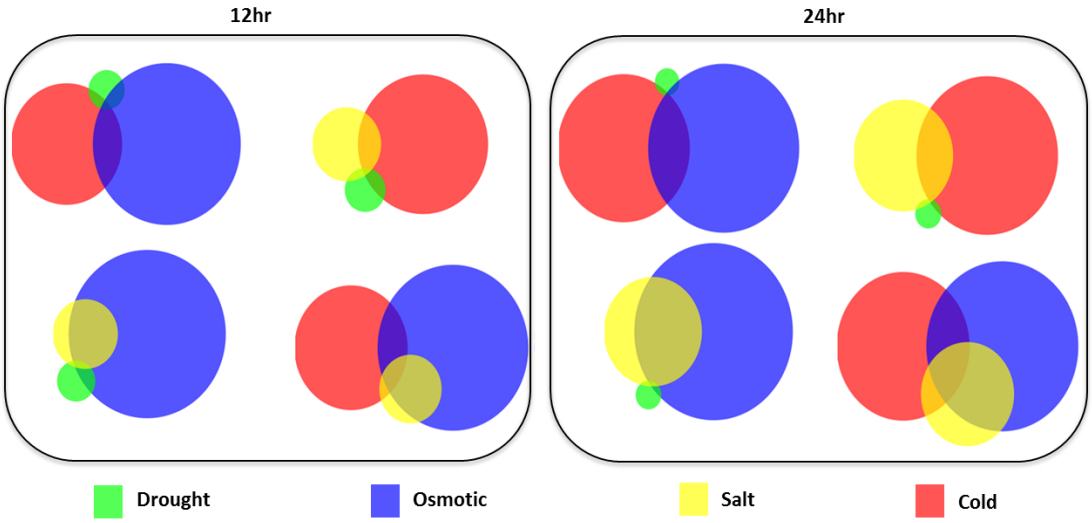


Figure 3.16: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 12hr and 24hr Shoots putative responsive sets. Each circle represents a CRE set.

Similar proportions were observed for CREs putative responsive to Cold, Osmotic, Salt and Drought stresses in roots tissues. In general the amount of Drought-responsive elements is greatly reduced after time point 1hr (see **Chapter 7.8**). CREs predicted to be responsive to Salt-stress 6hr displayed a clear majority in comparison with the other stresses (see **Figure 3.17**). Overall the results indicate that there is an overlapping set of CREs responsive to the abiotic stresses at time points 0.5hr and 1hr, which could act as convergence point for these abiotic signaling pathways. From time point 3hr Drought responsive elements are dramatically reduced in comparison with the other abiotic stresses, also the overlapping elements are reduced, indicating that drought responses do not crosstalk with the other analyzed stresses.

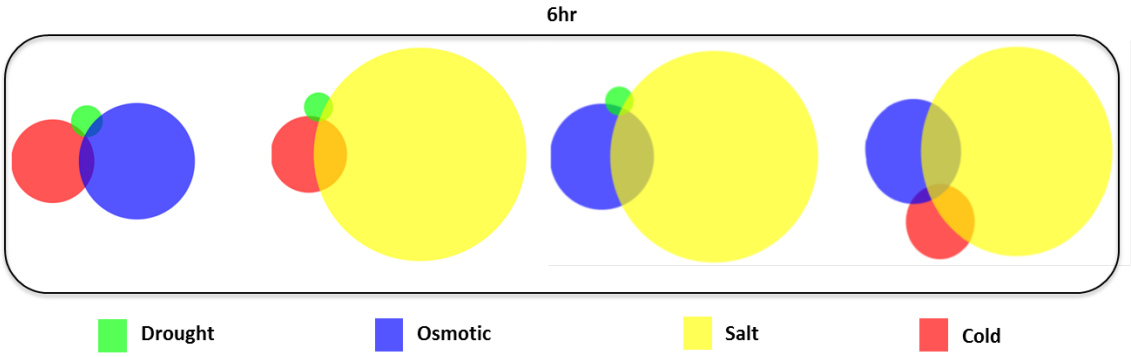


Figure 3.17: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 6hr Roots putative responsive sets. Each circle represents a CRE set.

3.3 Combinatorial *cis*-regulatory elements

One goal of the present study was to predict putatively functional combinatorial CREs. It has been previously shown that combinatorial elements occur with spatial constraints (spacer lengths, motif order and motif orientation) in different organisms (Yu et al. 2006; Singh 1998). Therefore, a new program described in **Chapter 2.8** was developed to predict combinatorial elements in *Arabidopsis thaliana.* The program uses input motifs and it searches combinations of such motifs within the *Arabidopsis thaliana* genome. The input motifs were predicted with the program MEME and are shown in the electronic appendix **E_3.3.1**. The program can predict motif combinations with and without spatial constraints. In order to test the effect of these constraints on element predictions (see **Chapters 3.3.1**), analyses of the spatial constraints described in **Chapters**, **3.3.2**, **3.3.3** and **3.3.4** were performed. Furthermore, it was tested in **Chapter 3.3.5** if the predicted combinatorial elements display characteristic distances to the TSS.

3.3.1 Spatial constraints

Combinatorial elements were predicted in the present study by applying spatial constraints. Such constraints include element order within the promoters (right and left positions), spacer lengths between motifs forming a combinatorial element and the elemnt relative orientation to the promoters (5’ to 3’ or 3’ to 5’ i.e. -> or <-). To test the applicability of such constraints on element predictions, two sets were predicted and used for expression analysis: one where combinatorial elements occur

without the above mentioned constraints and another set with elements displaying all spatial constraints. These two sets were directly compared to determine which one contains more putatively functional elements after expression analysis. Thus, the importance of applying these spatial constraints was assessed in order to predict functional combinatorial elements.

Following the pipeline described in **Chapter 2.8,** combinatorial elements putatively responsive to 18 different biotic and abiotic stresses were predicted (see **Table 3.18**). A total of 788 functional elements were predicted when applying spatial constraints. On the other hand, only 16 functional combinatorial elements without application of spatial constraints were predicted. By assessing the number of functional combinatorial elements predicted for each stress, it is observed that this number is always lower for functional elements without spatial constraints. The number of predicted functional combinatorial elements is greatly reduced (in comparison with combinatorial elements displaying spatial constraints) for the stresses EFTu60min, Flg22 1h and 4h, or even completely absent for the rest of stresses. Finally, no functional combinatorial elements were predicted associated to the abiotic stresses Pb 25ppm leaves, Zn-oversupply 8h roots and Zn-resupply shoots 8h vs. deficiency.

Table 3.18: Number functional combinatorial elements predicted with and without spatial constraints after expression analysis.

Stress	**Elements with spatial constraints**	**Elements without spatial constraints**
Chitooctaose	*30*	*0*
EF-Tu 30min	*68*	*0*
EF-Tu 60min	*76*	*11*
Flg22 1h	*54*	*4*
Flg22 4h	*116*	*1*
Pb 25ppm leaves	*0*	*0*
Pb 25ppm roots	*31*	*0*
Pb 50ppm leaves	*27*	*0*
Pb 50ppm roots	*77*	*0*
Zn-deficiency roots	*26*	*0*
Zn-deficiency shoots	*81*	*0*
Zn-oversupply 2h roots	*7*	*0*
Zn-oversupply 8h roots	*0*	*0*
Zn-oversupply 8h shoots	*1*	*0*
Zn-resupply roots 2h vs. def	*71*	*0*
Zn-resupply roots 2h vs. suf	*67*	*0*
Zn-resupply shoots 8h vs. def	*0*	*0*
Zn-resupply shoots 8h vs. suf	*56*	*0*
Total	*788*	*16*

Taken together, these results indicate that spatial constraints are important for the prediction of functional combinatorial element by expression analysis. The low number of combinatorial elements predicted when no spatial constraints are applied can be explained by the fact that elements without constraints occur in a very high number of gene promoters and therefore would not display statistical significant values after expression analysis. But this does not mean that there are no functional combinatorial elements without constraints, they are just not detectable using this method. Specific spatial constraints were further analyzed and the results are presented in the following chapters.

3.3.2 Element spacer lengths

Having observed that application of spatial constraints is important for combinatorial element prediction (see last chapter), the spacer length (distance between two motifs forming a combinatorial element) was further assessed. Each predicted combinatorial element for the different stresses can have different spacer lengths. Thus, the sets of combinatorial elements with spatial constraints were analyzed to determine if certain spacer lengths are overrepresented among the predicted elements. For this purpose, the frequency of the observed lengths among the predicted combinatorial element sets was determined. With that information, graphs were generated for each set where the number of elements having a certain spacer length within a window size of 50bp was shown (see **Chapter 2.8.2**). **Figure 3.18** displays the number of combinatorial elements putatively responsive to the stress Flg22 1h according to the observed spacer length frequencies. It was observed in this set that the majority of elements have a spacer length between 0 and 50bp followed by elements with lengths between 151 and 150.

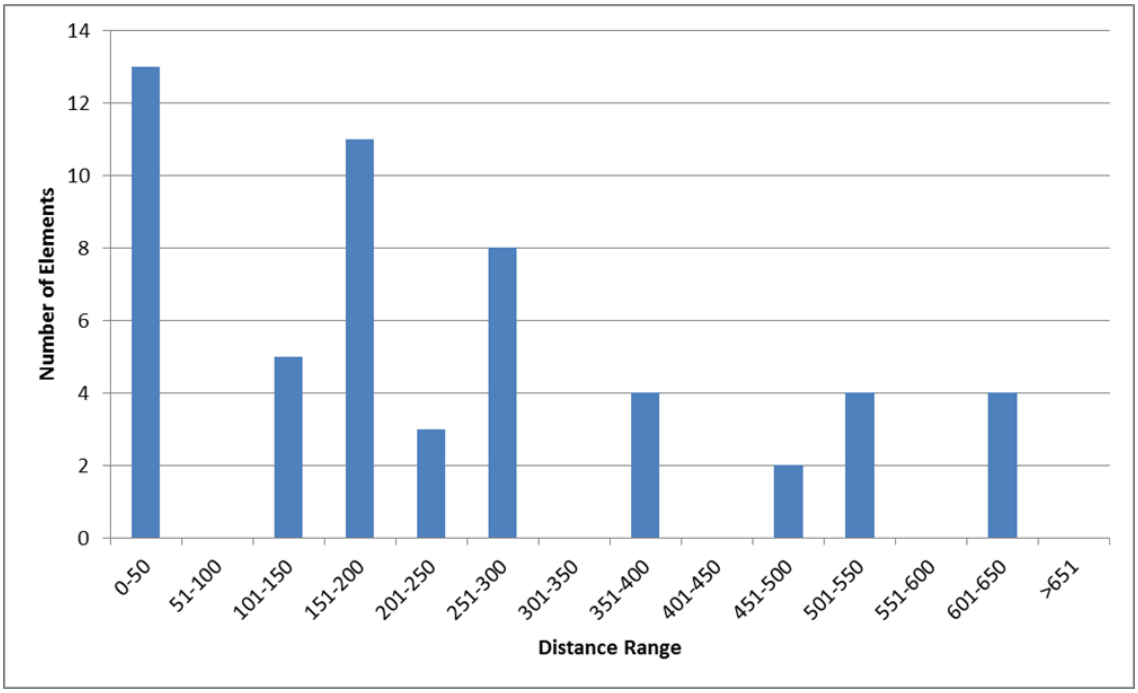


Figure 3.18: Spacer length frequencies of putative Flg22 1h responsive combinatorial elements in ranges of 50bp.

It was also of interest to assess if the observed spacer length frequencies occur more often than randomly expected. Therefore a random model was generated which

indicated the theoretical number of combinatorial elements randomly expected within a given set. Then, such expected values where compared with the observed ones in order to determine if there are statistical significant differences (see **Chapter 2.8.2**). **Figure 3.19** shows such comparison for combinatorial elements putatively responsive to Flg22 1h. Combinatorial elements having a spacer length between 0 and 50bp show the highest deviation from the randomly expected elements in this set. A frequency of 13 combinatorial elements with a spacer length of 0-50bp is not very likely to occur implying that this frequency strongly deviates from the expected value. In addition, there is a statistical significant higher than randomly expected number of elements with spacer lengths lying within the ranges 151-200. Furthermore, certain ranges display no elements for the set. Notably, the absence of elements with a spacer length of 51-100bp also shows a low probability. Together the results indicate that there are combinatorial elements in this set showing higher and lower than randomly expected spacer length distributions.

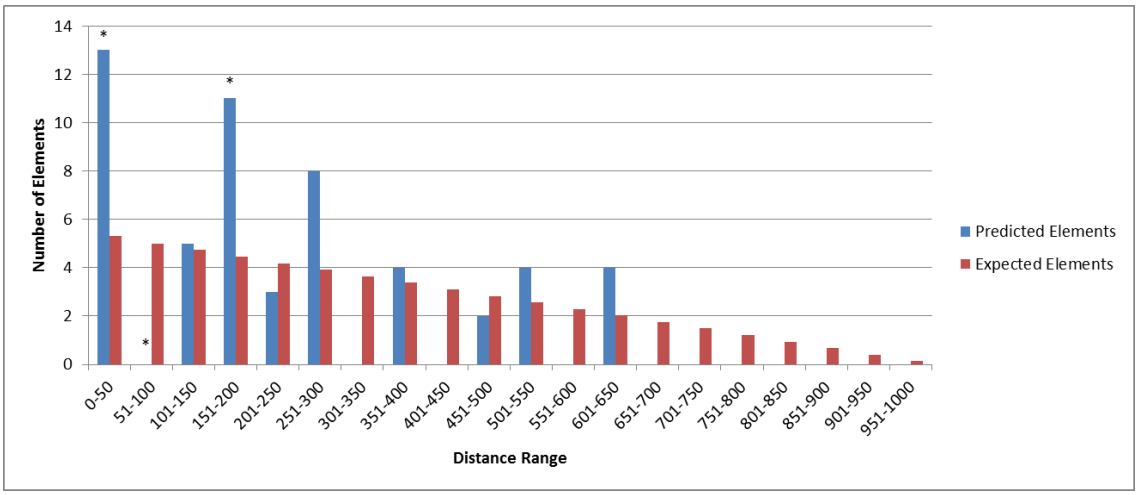


Figure 3.19: Distribution of combinatorial elements putatively responsive to the stress Flg22 1h according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

Characteristic spacer lengths were also observed among other predicted combinatorial element sets. **Figure 3.20** shows spacer lengths frequencies of elements putatively responsive to Flg22 4h. In this set, combinatorial elements with spacer lengths between 51-100bp have the highest occurrence and also a very low probability of

being observed. This length contrasts with the distances observed for the Flg22 1h set, where no element was observed with that spacer length. In addition, lengths which were a majority in the Flg22 1h set (0-50 and 151-200) do not display frequencies higher than expected. Other combinatorial elements among the predicted sets also displayed short spacer lengths (<100bp) with a frequency higher than expected, namely elements putatively responsive to Chitooctaose, EF-Tu 30min, Zn-resupplied roots 2h vs. sufficient Zn and Zn-resupplied shoots 8h vs. sufficient Zn (see **Chapter 7.4** for figures displaying spacer lengths).

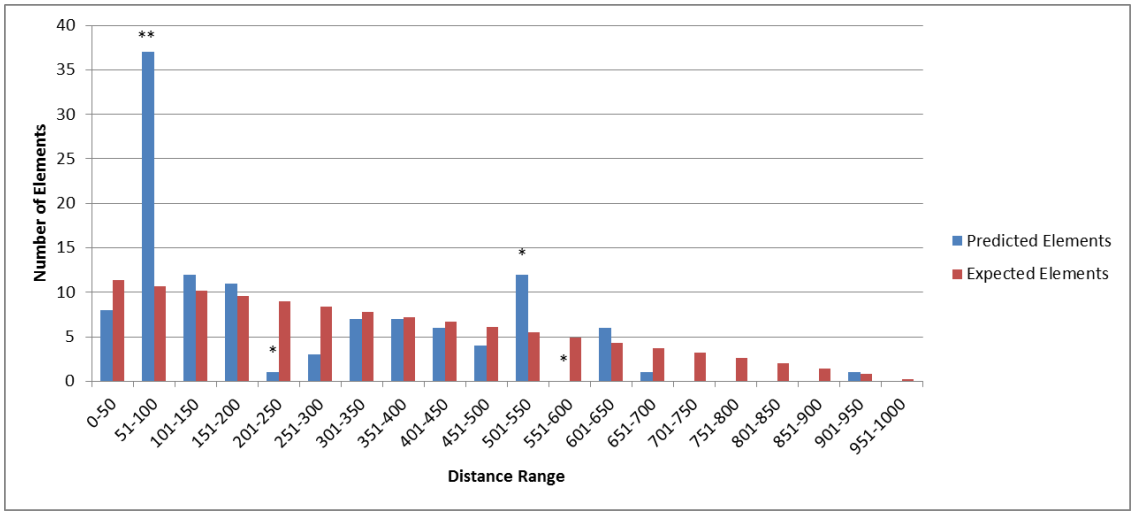


Figure 3.20: Distribution of combinatorial elements putatively responsive to the stress Flg22 4h according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

But not only short spacer lengths were the most frequent among all sets. Certain stresses displayed only longer spacer lengths with a statistically significant higher frequency than expected. This is the case for the set of combinatorial elements putatively responsive to the stress EF-Tu 60min (see **Figure 3.21**), where a high number of elements have a spacer length lying in a range of 501-550 and 651-700bp, both with a very low probability of occurrence. Only longer spacer lengths significantly higher than expected were also observed for combinatorial elements putatively responsive to the stresses Pb 25ppm and 50ppmroots, and Pb 50ppm leaves (see **Chapter 7.4** for figures displaying spacer lengths). These results indicate that certain spacer lengths seem to be stress-specific and that they occur with a statistically significant frequency.

Nevertheless, combinatorial elements associated to the stresses Zn-resupplied roots 2h vs. deficient Zn and Zn-deficient roots and shoots, showed no statically significant frequencies of spacer lengths.

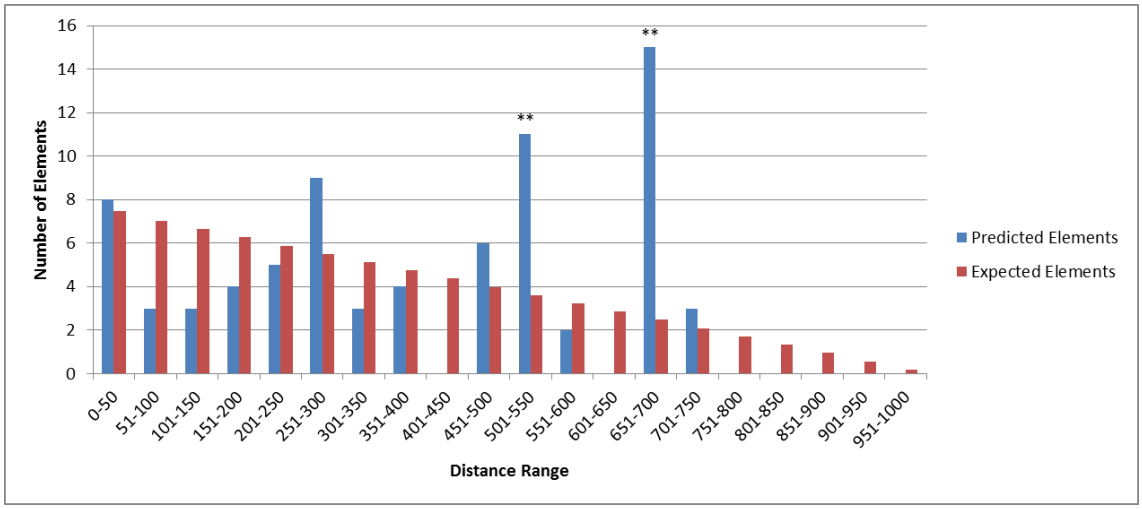


Figure 3.21: Distribution of combinatorial elements putatively responsive to the stress EF-Tu 60min according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

Finally it was assessed if there is an overall spacer length frequency among all predicted combinatorial elements occurring with a statistically significant probability regardless of a specific stress. For this purpose, the frequencies of the observed lengths among all predicted combinatorial elements were determined. Such values were also compared to randomly expected ones to generate **Figure 3.22**. The figure shows that the majority of combinatorial elements have a spacer length $\leq$ 100bp with a very low probability of occurring, which means that, overall, short spacer length distances are more common than longer lengths in the predicted sets. Also, combinatorial elements with long spacer lengths (501-550 and 651-700bp) also show statistically significant deviations from the expected values. In addition, several frequencies showed significantly less than expected combinatorial elements. In general, it is clear that short distances are a majority in the predicted combinatorial elements sets.

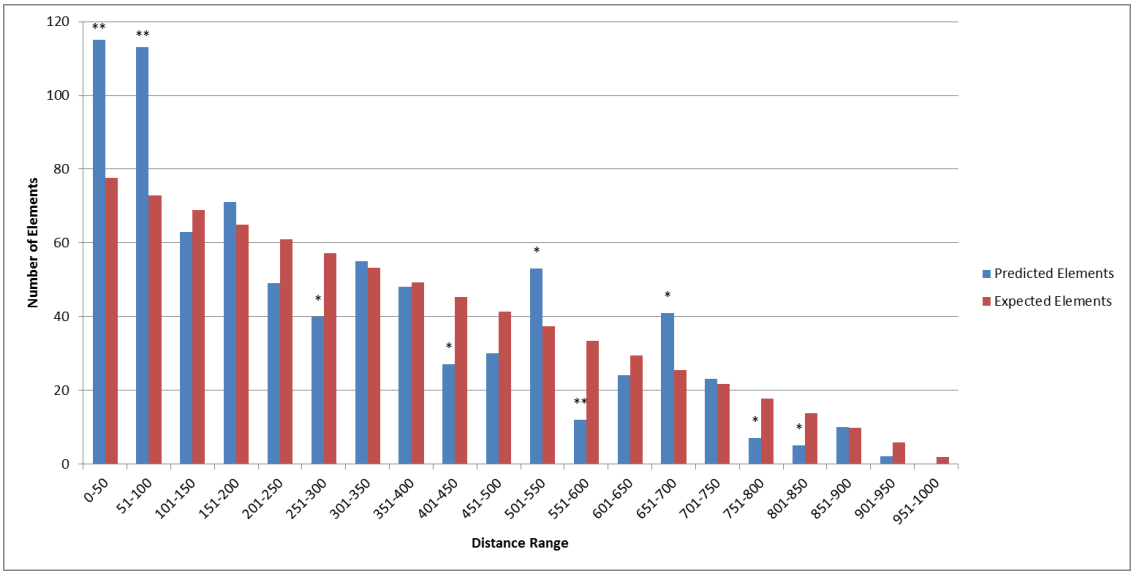


Figure 3.22: Distribution of all predicted combinatorial elements according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability ≤0.01 and ** a probability ≤0.001 of observing exactly the number of predicted elements given the expected ones.

3.3.3 Element orientation

Another spatial constraint in a combinatorial element is the relative orientation to the promoters (5' to 3' or 3' to 5' i.e. -> or <-) of the motifs forming the element. Motifs comprising combinatorial elements with spatial constraints predicted in the present study have specific relative orientations (initially defined by MEME). Thus, it was possible that elements in the predicted sets also show characteristic combinations of relative orientations. In order to test this, the frequency of such orientation combinations was determined for each predicted set. All possible four combinations of orientations, i.e. ->->, -><-, <--> and <-<-, were estimated for each predicted set of combinatorial elements. For visualization, the frequencies were used to construct graphs showing the percentage of a given orientation combination within a set. **Figure 3.23** shows the orientation frequency among Flg22 1h responsive elements. The orientation combination with the highest frequency is <-->, however with a 31%. Orientations -><- and ->-> are present both with a frequency of 24% and 30% and ->-> with 22%. There is however no clear major orientation preference in the set.

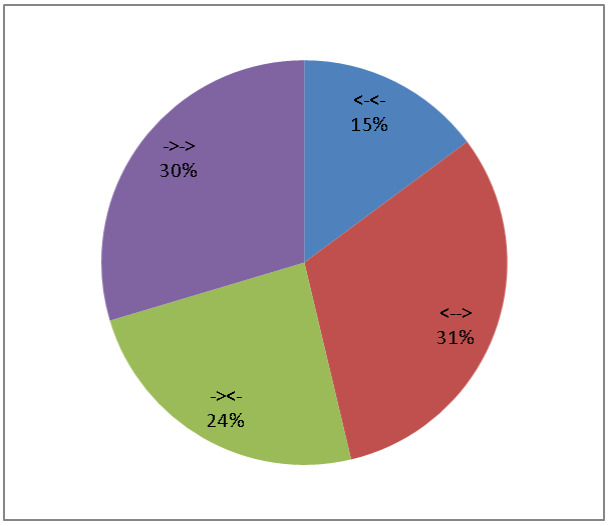


Figure 3.23: Orientation frequency of putatively Flg22 1h responsive combinatorial elements.

The frequency of orientations for the remaining sets was also calculated, graphs are presented in **Chapter 7.5, Table 3.19** summarizes the data. Chitooctaose, EF-Tu 30min and Zn-deficiency roots responsive combinatorial elements seem to have a preference for the orientation combination ->->. Combinatorial elements putatively responsive to the abiotic stresses Pb 25ppm roots, Pb-50ppm leaves and Zn-oversupply 2h roots show <--> as the most frequent orientation combination. The remaining sets show no clear orientation preferences, as can be seen in **Table 3.19**. Overall, combinatorial elements show a small frequency (17%) of elements with the orientation <-<-, however a general preferred orientation combination is not observed. These results suggest that, although single predicted combinatorial elements have specific orientation combinations, there is no general clear preference for the predicted combinatorial element sets.

Table 3.19: Frequency of orientation combinations among predicted combinatorial element sets.

Stress	Percentage			
	->->	**-><-**	**<-->**	**<-<-**
Chitooctaose	40	27	17	16
EF-Tu 30min	57	27	13	3
EF-Tu 60min	28	13	31	28
Flg22 1h	30	24	31	15
Flg22 4h	22	23	32	23
Pb 25ppm roots	19	6	65	10
Pb 50ppm leaves	38	12	47	3
Pb 50ppm roots	29	16	36	19
Zn-deficiency roots	52	19	22	7
Zn-deficiency shoots	28	38	16	18
Zn-oversupply 2h roots	19	12	50	19
Zn-resupply roots 2h vs. def	27	20	38	15
Zn-resupply roots 2h vs. suf	27	28	29	16
Zn-resupply shoots 8h vs. suf	10	34	30	26
Total	29	24	30	17

3.3.4 Element order

Predicted combinatorial elements have a further spatial constraint, namely the order of the motifs forming the element. A combinatorial element is formed by two motifs separated by a spacer length. Such two motifs can have either constant positions, i.e. one motif is always on the right and the other on the left, or variable positions within promoters. In order to test how important this spatial constraint is, one further combinatorial element set with variable motif order was generated and compared to the set with constant motif order.

The pipeline described in **Chapter 2.8** was used to predict a set of combinatorial elements with different motif order within promoters. The set was compared with combinatorial elements having constant motif order (see **Table 3.20**). It is observed that the total number of combinatorial elements with different motif order is higher than elements with constant order. This is especially the case for combinatorial elements responsive to Flg22 1h and 4h, where the number of elements almost doubles for elements with different motif order. On the other hand, a similar number

of combinatorial elements predicted to be responsive to abiotic stresses was predicted for both sets (with constant and different positions).

Table 3.20: Comparison of combinatorial elements number having constant and different motif order.

Stress	Number of combinatorial elements		Number of different combinatorial elements	
	Different motif order	Same motif order	Different motif order	Same motif order
Chitooctaose	47	*30*	13	*9*
EF-Tu 30min	83	*68*	25	*19*
EF-Tu 60min	113	*76*	22	*18*
Flg22 1h	102	*54*	22	*16*
Flg22 4h	207	*116*	23	*16*
Pb 25ppm leaves	0	*0*	0	*0*
Pb 25ppm roots	31	*31*	16	*16*
Pb 50ppm leaves	27	*27*	10	*10*
Pb 50ppm roots	83	*77*	47	*44*
Zn-deficiency roots	32	*26*	21	*17*
Zn-deficiency shoots	89	*81*	39	*34*
Zn-oversupply 2h roots	7	*7*	5	*5*
Zn-oversupply 8h roots	0	*0*	0	*0*
Zn-oversupply 8h shoots	1	*1*	1	*1*
Zn-resupply roots 2h vs. def	77	*71*	38	*29*
Zn-resupply roots 2h vs. suf	70	*67*	26	*25*
Zn-resupply shoots 8h vs. def	0	*0*	0	*0*
Zn-resupply shoots 8h vs. suf	63	*56*	31	*27*
Total	1032	*788*	339	*286*

The similarities among the combinatorial elements in the predicted sets were assessed as described in **Chapter 2.8.1**. By determining these similarities, the analysis allowed then the identification of the number of different combinatorial elements among each set. This served to measure if the analysis for elements with different motif order also yields a higher variety of combinatorial elements when compared to elements with constant order. The comparison is also displayed on **Table 3.20**. The differences in the different number of predicted elements between equal and different motif order sets are very low when the elements similarity is considered. For example, although the number of combinatorial elements responsive to Flg22 1h and 4h was almost doubled in the different order set, the number of different elements is similar. This means that,

although there are a higher number of predicted elements with different motif order, this increase has no effect on the element variety, i.e. elements similar to the ones already predicted are being found. The differences between both sets were further analyzed by assessing the number of predicted and expected elements in the sets. The comparison for elements predicted to be responsive to Flg22 1hr is shown on **Figure 3.24**, where it is possible to observe that the increase of combinatorial elements having different motif order is mainly due to increases in the number of randomly expected elements. This is very clear by observing the increase of elements with a spacer length between 51 and 100bp. Together the results indicate that predicting combinatorial elements with different motif order within promoters increases the number of randomly expected combinatorial elements and does not affect combinatorial element variety.

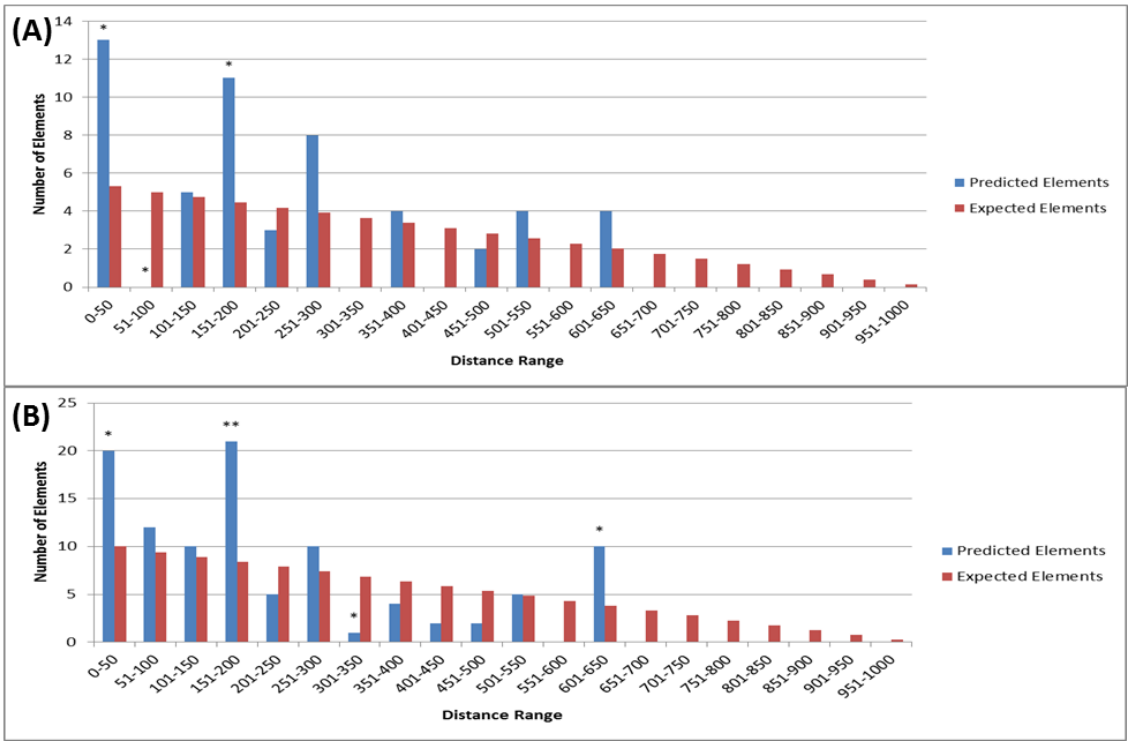


Figure 3.24: Comparison between putatively Flg22 1h responsive combinatorial elements having different (A) and equal (B) motif order within promoters. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

3.3.5 Element distance to TSS

The last spatial constraint assessed for the predicted combinatorial elements was the distance to the TSS. The goal was to determine if there is any characteristic distance occurring in the predicted combinatorial elements with spatial constraints. As described in **Chapter 2.8.3**, distance to the TSS of combinatorial elements can be measured in two ways: distance from the nearest motif or from the farthest motif forming the combinatorial element (see **Figure 2.14** in page49). Elements occur within gene promoters with a fixed spacer length but with different positions among the promoters (see **Figure 3.25**). This means that a single combinatorial element shows different distances to the TSS. Thus, for each gene promoter where predicted combinatorial elements occur, the distance to the TSS from both the nearest and farthest motifs was determined. These distances were calculated for ranges of 50 nucleotides by assessing how many genes have elements with distances to the TSS between 0-50, 51-100 until 951-1000bp. By generating a random model (see **Chapter 2.8.3**), the random expectancies of the distances in every range, as well as the probabilities of a frequency being observed given the expected values were calculated.

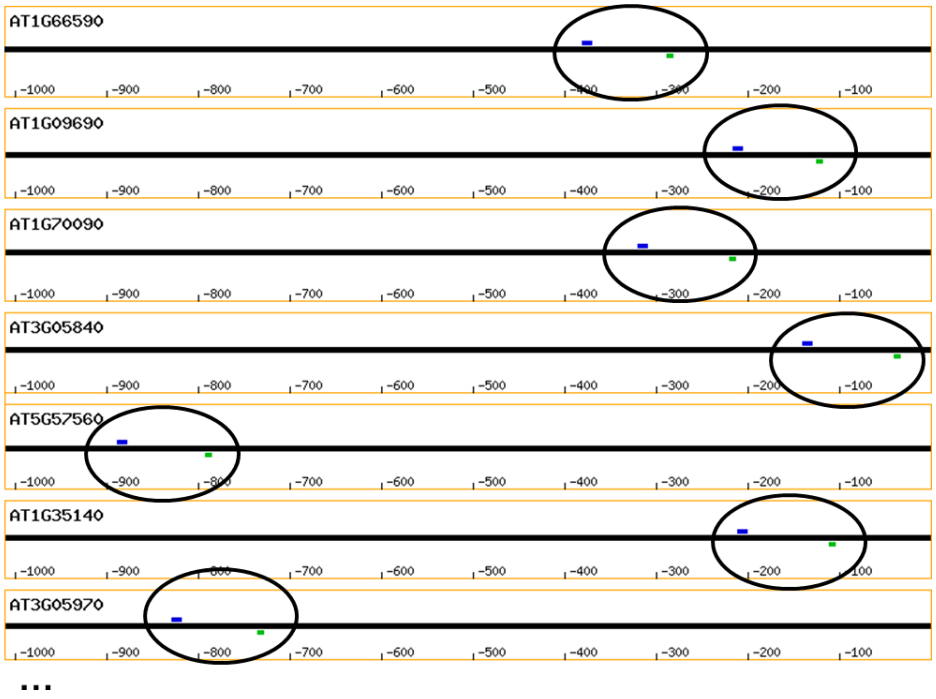


Figure 3.25: Combinatorial element (highlighted in circles) predicted to be responsive to EF-Tu 60min. Each gene promoter where the element occurs is shown.

Frequency of distances to the TSS from the nearest and farthest motifs forming combinatorial elements putatively responsive to Flg22 1h are shown on **Figure 3.26**. By assessing these frequencies with their corresponding probabilities, the distances to the TSS from the nearest and farthest motifs show several ranges with significant probabilities. For the farthest motifs in combinatorial elements, distances to the TSS ≤ 550bp display lower than expected frequencies, whereas distances ≥ 551bp occur with a frequency higher than expected. The distances to the TSS from the farthest motifs show also several lengths (≤ 500bp) with a frequency higher than randomly expected. On the other hand, distances ≥ 551 show frequencies with lower than expected probabilities. Thus, these results suggest that, although no characteristic distance to the TSS was observed, the distance distributions to the TSS do not seem to follow a random distribution.

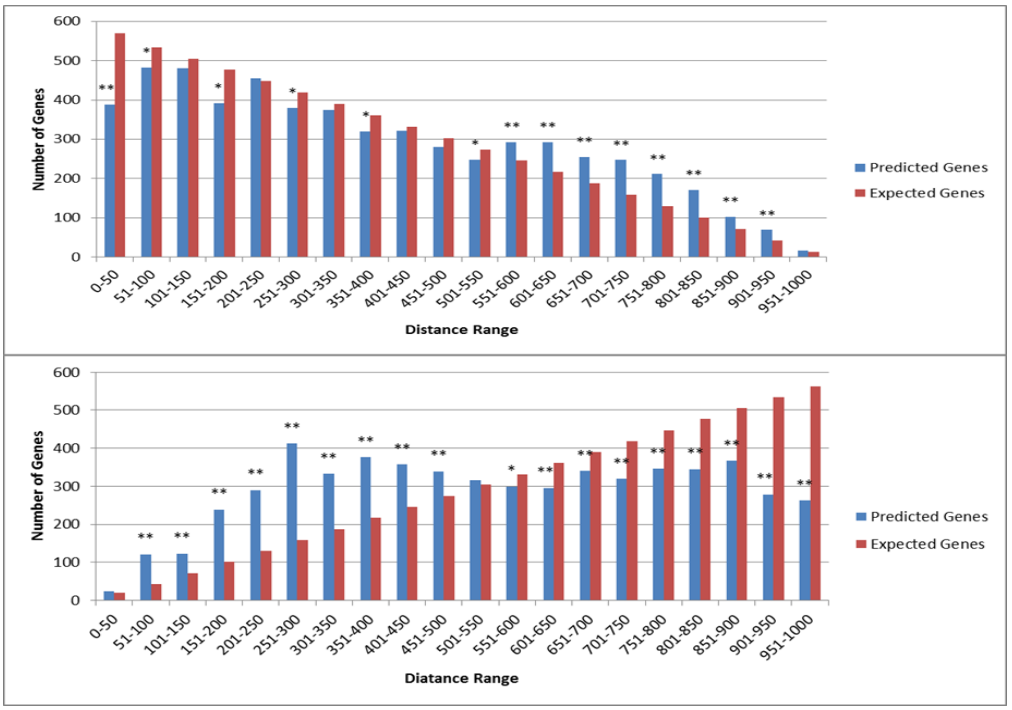


Figure 3.26: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Flg22 1h. An * indicates a probability ≤0.01 and ** a probability ≤0.001 of observing exactly the number of predicted elements given the expected ones.

Frequencies of distances to the TSS from combinatorial elements putatively responsive to the stress EF-Tu 30min are shown in **Figure 3.27**. The majority of the nearest motifs have a distance which lies within a range of 0-50bp with a frequency higher than expected. However, other distances are also observed with a low probability of

occurring randomly. Also the distances to the TSS from the farthest motifs, rather than showing single characteristic peaks, they display an overall distribution that does not seem to occur by chance. Similar distance distributions of distances to the TSS were also observed for other predicted combinatorial elements responsive to the analyzed stresses (see **Chapter 7.6**). These results indicate that overall, distances do not seem to occur with a randomly expected frequency within the predicted sets, although no characteristic single distance to the TSS was observed.

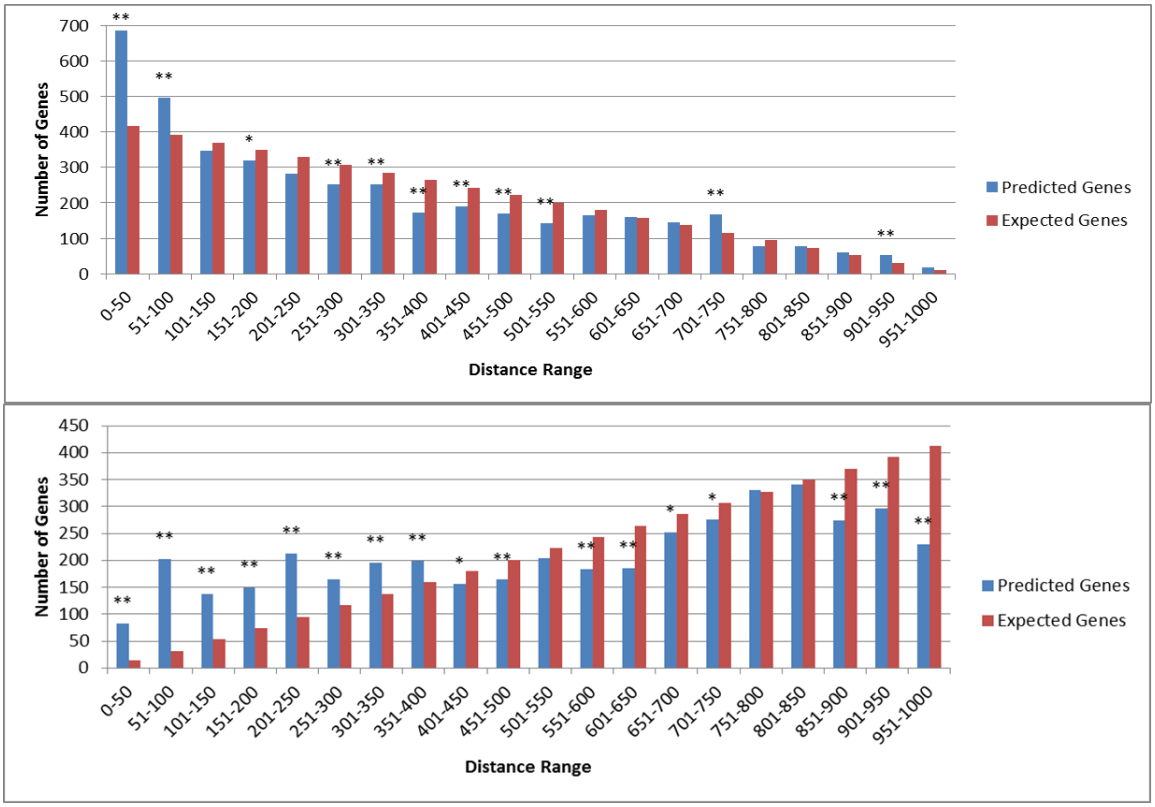


Figure 3.27: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to EF-Tu 30min. An * indicates a probability ≤0.01 and ** a probability ≤0.001 of observing exactly the number of predicted elements given the expected ones.

4 Discussion

4.1 *In silico* expression analysis as a tool for *cis*-regulatory element prediction

Motif finding programs predict a large number of possible sequences as CREs. The time and resources needed to experimentally validate such a large number of sequences makes the task very difficult to accomplish. There is then a clear need to pre-select sequences for experimental analysis which have a higher probability of being functional CREs. Approaches already reported to solve this problem will be discussed later. However, solutions specifically for plants seem to be very limited. As an approach to solve this, a new method called the *in silico* expression analysis for the prediction of plant CREs was developed. The analysis correlates motif or sequence occurrences within *Arabidopsis* gene promoters with gene expression data from the PathoPlant database in order to predict putatively functional CREs.

The power of microarrays for gene expression analysis in plants has been extensively studied (van Hal et al. 2000; Aharoni and Vorst 2002; Cushman and Bohnert 2000; Maruyama et al. 2004). Furthermore, similar analyses correlating expression for CRE detection have been reported for other organisms. (Bussemaker et al. 2001) developed an algorithm to find CREs in *Sacharomyces cerevisae* based on a model where upstream motif occurrences contribute to the expression of a given gene. Similar to the present study, gene promoter sequences and expression information are used as input for CRE detection which is performed by selecting the most statistically significant motifs which could control gene expression (Bussemaker et al. 2001). Also using genome and expression data of *Sacharomyces cerevisae,* (Caselle et al. 2002) developed a computational method for CRE detection. The method is very similar to the one implemented in the present study: the average expression in yeast, when shifting from fermentation to respiration of genes containing a given sequence within their promoters is calculated. When this average expression is significantly higher or lower than the overall expression of all genes, the shared sequence is predicted to be a putatively functional CRE. Also by correlating motif occurrences and gene expression

(Janaki and Joshi 2004) was able to find possible CREs responsive to diurnal rhythms in *Arabidopsis thaliana*. Thus, the similarities between the *in silico* expression analysis developed in this study and the above mentioned methods suggest that the analysis is functional and a very useful tool for prediction of CREs. An advantage of the *in silico* expression analysis over the other mentioned methods is that it allows testing several conditions or stresses for each CRE candidate, which serves as a measure of not only functionality but also specificity.

For plants, a new tool to perform the *in silico* expression analysis online has been developed in the course of the present study, the tool is freely available as a web service at http://www.pathoplant.de/expression_analysis.php. It offers the possibility to enter a short DNA sequence to be searched in *Arabidopsis thaliana* gene promoters for the identification of genes putatively regulated by the sequence. Furthermore, it can be selected if genes putatively regulated post-transcriptionally by small RNAs and microRNAs should be excluded from the calculation, which may improve the quality of the predictions by exclusively analyzing transcriptionally-regulated genes. Genes being putatively regulated by smallRNAs and microRNAs were identified using tools developed for the AthaMap database (Bülow et al. 2009) (Bülow 2012). As a result of the *in silico* expression analysis, the stresses for which the submitted sequence is most probable responsive are given within a ranked list including statistical significance information (see **Chapter 3.1.5**). Thus, the web tool serves as a valuable resource for the plant science community to *in silico* evaluate the potential functionality of a given sequence as a CRE.

Another web tool called *cis*-elements was developed in the course of this study and is also planned to be a freely accessible tool at http://www.pathoplant.de/. It is the reverse of the *in silico* expression analysis test, since in this case a certain stress is selected in order to retrieve possible CREs associated to it. Such CREs are identified from the analysis, where all possible DNA 10mer combinations were generated and tested with the *in silico* expression analysis. For the online tool, the possibility of selecting DNA 9mers and 8mers was also implemented. The analysis has several advantages over the classical methods of finding overrepresented motifs in up-regulated genes. First, the fact that all possible DNA 10mers, 9mers and 8mers

combinations are tested does not rely in a program to find overrepresented motifs. This is particularly useful, since such programs use promoter regions of genes with similar expression profiles as input and it has been shown that in many cases not all such regions display sequence similarity (Blanco et al. 2006), making the finding of functional CREs difficult to motif-finding programs. The tool offers as an additional option the possibility of selecting the specificity of the predicted CREs by determining which type of stresses should also display significant values. In that way, CREs putatively specific to one or more stress types (abiotic, biotic, fungal, other) can be predicted. This feature can also be used to refine the results, since the fact of observing a sequence with a high number of significant values for a similar stress type confirms the results from several independent experiments. The strong importance of these independent confirmations in gene expression analysis using microarray data has been stated by (Firestein and Pisetsky 2002). Thus, the advantages of the *cis*-elements tool make the web service a useful on line resource for specific CRE detection in plants.

DNA microarrays have nevertheless limitations, as noted by Shendure (2008). Sequences that are very similar will show cross-hybridization making the analysis of related sequences problematical. Circumventing such problems, massive parallel sequencing or Next Generation Sequencing (NGS) technologies are available for transcriptome analysis (Reis-Filho 2009). The principles of the analyses performed in this study are also applicable with NGS data. Thus, it is expected that by using such data the predictions performed by tools like the *in silico* expression analysis will be improved.

4.1.1 Proof of concept

The *in silico* expression analysis developed in the present study was validated by assessing expression values obtained when known CREs are used as input and by analyzing synthetic CREs coming from an experimental approach. The known DRE element with the sequence TACCGACAT, reported by (Yamaguchi-Shinozaki and Shinozaki 1994) to be responsive to Dehydration-, High-Salt-, and Low-Temperature was analyzed with the *in silico* expression analysis. The most significant stresses associated with this sequence turned out to be cold stresses (see **Table 3.1** in page 52).

Other analyses also confirmed the responsiveness of this DRE element to cold stresses. In *Arabidopsis thaliana* (Kim et al. 2002) a transcription factor was reported to bind to the DRE sequence in response to cold and dehydration (Stockinger et al. 1997). In addition, the *in silico* expression analysis predictions correlate with the observation made by (Shinwari et al. 1998) that certain genes containing the DRE element where mainly induced by cold stresses in *Arabidopsis*. These experimental data confirm very well the observed results from the *in silico* expression analysis for this DRE.

The palindromic sequence ATGTCGACAT was reported by (Assunção et al. 2010) to be present in the promoter regions of zinc-deficiency responsive genes. This Zinc Deficiency Responsive Element (ZDRE) is very important for the *Arabidopsis* primary response to Zinc-Deficiency (Assunção et al. 2010). The responsiveness of the ZDRE towards Zinc-Deficiency is also observed by analyzing the expression values after the *in silico* expression analysis (see **Table 3.3** in page 53). Other known CREs also showed expected responses (see **Chapter 3.1.1**) which served as a validation of the newly developed expression analysis tool.

As mentioned before, the *in silico* expression analysis was validated by analyzing also synthetic CREs coming from a high throughput experimental approach. This approach used pep25-elicited parsley protoplasts transformed with a random library of sequences to isolate a very high number of synthetic *cis*-regulatory elements potentially responsive to fungal stresses. This huge set of synthetic CREs was subsequently validated using the *in silico* expression analysis tool. In a comparative analysis including elements isolated from untreated protoplasts as controls, the set of elements after fungal elicitation showed a higher proportion of putatively functional *cis*-regulatory elements for fungal stresses than the control (see **Chapter 3.1.2**). In conclusion, it was possible to observe a successful enrichment of specific elements that seemed to be induced upon fungal elicitation in the experimental screening. The specificity of the enriched synthetic elements was observed by assessing the differences in overall expression values calculated for both (enriched and control) samples. The elements in the control sample displayed indifferent responsiveness patterns to stresses, whereas the enriched sample showed significant specificity to fungal pathogens due to a depletion of unspecific elements. Furthermore, the

frequency of repetitions from the synthetic CREs also has an effect on the expression values by making the synthetic elements more specific towards fungal stresses (see **Figure 3.3** in page 58). This was also an expected result, since in the experimental approach the most frequent synthetic elements are the elements being more often actively transcribed, and in the enriched sample such elements were expected to be more responsive to fungal pathogens. Thus, the validation of the elements serves as a good proof-of-concept for both the *in silico* expression analysis using known CREs as well as experimentally pre-selected CRE sequences. The analysis has yielded promising fungal-responsive candidate *cis*-regulatory elements for further experimental analyses.

4.1.2 CREs predictions

Motifs predicted by the program MEME were used as input data for the *in silico* expression analysis in order to predict putatively functional CREs. This analysis yielded a mixture of novel and known CREs (see **Chapter 3.1.3**). These predicted motifs were further assessed with a STAMP analysis to determine motif diversity, which turned out to be low, since the 1014 predicted motifs were clustered into 261 different groups (see **Table 3.5** in page 59).

In order to further improve the predictions performed with the *in silico* expression analysis and to increase the diversity of the predicted motifs, several methods were implemented. The first step was to use input sequences generated independently from an algorithm for finding overrepresented motifs. Although algorithms like MEME have performed very well in benchmarking tests, the performance of such algorithms depends on the sequences used as input data (Hu et al. 2005), which requires a previous clustering of genes according to expression profiles. As an approach to overcome this limitation, all possible DNA 10mers (1,048,576 sequences) were used as input elements for the *in silico* expression analysis. This novel method served as an independent analysis from motif finding programs. As observed in the results obtained with input motifs predicted by MEME, the motifs predicted with the novel method are comprised of a mixture of novel and known CREs (see **Chapter 3.1.4**), which serves to validate the approach.

In order to select putatively functional as well as specific CREs, a very important criterion was introduced. For each predicted sequence the average expression of the genes containing such sequence within promoters was calculated upon different stresses. Thus, it was possible to assess which stresses and stress types (abiotic, biotic, fungi and others) showed statistically significant expression values. This is an approach to evaluate the specificity of the predicted sequences. Different stress signaling pathways crosstalk forming a very complex network where certain molecules are specific to certain stresses (Genoud and Métraux 1999). In addition, bioinformatic methods have been developed for the prediction of plant CREs and the assessment of possible pathway crosstalks (Priest et al. 2009). The analyses performed in this study aimed also at finding CREs specific to Flg22 and Drought stresses by using crosstalk information. In order to accomplish this, gene expression values upon these stresses of interest were assessed in the context of the other stresses that can also show significant p-values. By selecting sequences present in genes showing statistically significant expression values in only the stress type of interest, CREs specific to that stress type are expected to be predicted.

Another level of specificity was introduced by generating a similarity tree with the predicted biotic and abiotic sequences to compare them with each other. Comparative analyses of DNA have been largely used for CREs prediction (He et al. 2009; Pierstorff et al. 2006; Siddharthan et al. 2005). Phylogenetic footprinting compares DNA regions among different species in order to find similar and conserved CREs (Tagle et al. 1988). In this study, the similarity trees were used to exclude mixed clusters containing biotic and abiotic stress-responsive elements. Clusters of CREs were identified that harbor no elements responsive either to abiotic predicted sequences in the case of Flg22 or biotic predicted sequences for drought stress (see **Figure 3.5** in page 66). The goal of this analysis was to further improve the selection of CREs according to sequence specificity. This analysis produced two sets of elements putatively responsive to Flg22 and drought stresses.

The majority of Flg22 responsive sequences from this set contained the well-known pathogen responsive W-box domain (TTGACT/C). The W-box is the binding site of WRKY transcription factors involved in plant defense mechanisms (Ulker and Somssich

2004). The link between WRKY transcription factors and Flg22 perception has been shown experimentally (Dong et al. 2003). Further predicted CREs from this set also displayed similarities to other pathogen-related motifs (Strompen et al. 1998). Elements from the set predicted to be responsive to Drought-stresses also showed similarities to known CREs related to abiotic-stresses. CREs similar to Abscisic Acid Responsive elements (ABREs) (Choi et al. 2000; Shen et al. 1996) were found, such ABREs had been shown to be involved in Drought responses in *Arabidopsis* (Fujita et al. 2005; Uno et al. 2000). Overall, the analysis yielded a known but also previously unreported and putatively functional pathogen- and drought-related CREs. To increase the variety of predicted CREs, the number of predicted sequences can be easily increased in both sets, since for this analysis only the top 30 Flg22 and Drought sequences from all predicted sequences for such stresses were analyzed. Candidate sequences from both sets (Drought and Flg22) were chosen for experimental validation and are currently being tested at the plant genetics lab of the technical university of Braunschweig.

As a perspective, the analysis described here could be further developed. It is known that the binding sites of a transcription factor are commonly degenerated. The fact that only perfect matches of single sequences (from all possible DNA 10mers) were used within the analysis does not allow the finding of degenerated binding sites. A future approach to solve this limitation could be the creation of motifs from the DNA10 mers set. The generation of motifs can be accomplished by using one of the common algorithms for global alignment such as the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). From these alignments, it would be possible to generate frequency matrices, which serve as an accurate representation of a binding site. These motifs could then be used as input data for an adapted version of the *in silico* expression analysis in order to find degenerate binding sites. The length of the input sequences could also be varied to include DNA 9mers and 8mers, which have also been used as input sequences for the online analysis tools and are already available for analysis. It is also known that transcription factors bind DNA cooperatively (Singh 1998). In this study, a new program was developed that uses Position-Specific Soring Matrices in order to find synergistic combinations of DNA binding sites. This program could use matrices generated from DNA10mers, 9mers and 8mers as input

data, which may improve the predictions for CREs. Such a tool would also strongly benefit from the results regarding combinatorial control of gene expression described in **Chapter 4.3**.

4.2 Pathway crosstalks

In the present study, microarray expression data was used to associate predicted CREs to various stresses thereby identifying possible signaling pathway crosstalks. The CREs putative-responsiveness to many stresses was determined with the *in silico* expression analysis data and newly developed tools. DNA microarrays have been extensively used for the study of signaling pathway crosstalks in plants (Seki et al. 2002; Narusaka et al. 2004; Schenk et al. 2000). In a similar approach to the present study, putatively functional CREs were predicted which were further used to assess the specificity of biotic and abiotic stress-responses (Zou et al. 2011). Possible biotic signaling pathway crosstalks predicted in the present study are described in **Chapter 4.2.1** and abiotic responses are discussed in **Chapter 4.2.2**.

4.2.1 Crosstalk in biotic stresses

In the present study, CREs putatively responsive to Flg22 were also predicted to be responsive to the bacterial elicitor Harpin Z and EF-Tu, indicating possible signaling pathway crosstalks among these stresses (see **Table 3.13** and **Table 3.14** in page 81). The common regulation observed for Flg22 and Harpin Z stresses correlates with previous studies. (Mészáros et al. 2006) demonstrated that a MAP kinase (MPK4) is activated by both Flg22 and Harpin Z elicitors, suggesting common responses for these stresses. In addition, further MAPKs have been identified as being targets of Flg22, Harpin and EF-Tu (Colcombet and Hirt 2008). Also consistent with the results observed in the present study, Flg22 and EF-Tu have been shown to activate similar gene sets and signaling events in plant defense responses (Zipfel et al. 2006). Interestingly, it was observed that crosstalks occur with stresses at similar time-points. Flg22, Harpin Z and EF-Tu were predicted to crosstalk at time point 1hr and Flg22 and Harpin Z also at time point 4hr (see **Table 3.13** and **Table 3.14** in page 81). This observation suggests that different signaling pathways are activated depending on the stresses and their time-

points. Correlating with these observations, (Swindell 2006) pointed out that gene expression responses were time-dependent. Furthermore, time-point specificity was also observed for combinatorial elements predicted in the present study, which, again, suggests that plant-defense responses are not only stress but also time-point specific.

Possible crosstalks between biotic and abiotic stresses were also observed. CREs putatively responsive to Chitooctaose were also predicted to be responsive to Salt stress (see **Table 3.15** in page 82). Correlating with these observations, *Arabidopsis thaliana* salt-stress responsive genes have been shown to be induced in biotic stresses (Ma et al. 2006). Furthermore, *Arabidopsis* genes have been also demonstrated to be responsive to salinity and chitin (Debnath et al. 2011). Very specific responses where observed for the CREs predicted to be responsive to Zn and Pb (see **Table 3.16** and **Table 3.17** in page 83), which indicates that responses to these stresses do not seem to crosstalk with other stress responses. Thus, the *in silico* expression analysis is presented as a valuable tool for the identification of possible pathway signaling crosstalks. This information can be used to predict CREs putatively responsive to a wide range of stresses or on the contrary, to identify CREs with a very stress-specific responsiveness.

4.2.2 Abiotic stresses regulation

In the present study CREs putatively responsive to Cold, Drought, Osmotic and Salt stresses at 6 different time-points were analyzed in order to determine how the elements overlap for each stress. Cold, Osmotic and Salt stresses were further analyzed because they were almost always present in the crosstalk analyses performed in this study. It has been suggested that abiotic stresses pose a higher danger to plants than biotic stresses (Fujita et al. 2006), which could explain the observation that these abiotic stresses are present in various crosstalk analyses. Drought stresses were also included in the analyses because it has been shown that these abiotic stresses have common convergence points which lead to pathway signaling crosstalks (Chinnusamy et al. 2004; Shinozaki and Yamaguchi-Shinozaki 2000). The fact that no crosstalk was observed between biotic and Drought stresses could be explained by the observation

that a requirement for successful pathogen attack in nature is humidity (Fujita et al. 2006).

The abiotic response was assessed in the present study by determining how many of the CREs putatively responsive to Cold, Drought, Osmotic and Salt stresses overlap. The analysis identified how many CREs are responsive to one or several stresses, which would indicate possible signaling convergence points (see **Chapter 3.2.2**). It was observed that for early time points (30min and 1hr) the Drought elements are overrepresented in comparison with the other abiotic stresses. The high number of overlapping CREs indicates possible crosstalks between these stresses. This observation is consistent with known crosstalks between these abiotic stresses (Chinnusamy et al. 2004; Shinozaki and Yamaguchi-Shinozaki 2000). In addition it has been reported that certain genes are strongly induced by salt, osmotic and drought stresses at early time-points (Ma et al. 2006), indicating a common regulation mechanism for early abiotic-responses. Furthermore, (Seki et al. 2002) reported that drought and high-salinity stresses display a higher degree of crosstalk than cold and high-salinity stresses. This correlates with observations from this study by analyzing the CREs for roots and time-point 1hr, where the number of overlapping CREs between salt and osmotic stresses is much higher than for cold and salt stresses. From time-point 3hr overlapping, as well as total drought CREs, are dramatically reduced in comparison with the other stresses and earlier time-points. This indicates that drought-responsive CREs are more specific at late time-points. On the other hand osmotic, salt and cold responsive CREs still display overlapping elements. Correlating with these observations, (Ma et al. 2006) reported the existence of genes being strongly up-regulated under salt and osmotic stresses after time-point 3hr, suggesting that crosstalks occur for these stresses at late time-points. The information gathered with this analysis can be used to either select CREs displaying a specific response to one abiotic stress, or to select elements with a wider responsiveness-spectrum towards abiotic stresses.

4.3 Combinatorial control in *A.thaliana*

In this study a new program was developed for the prediction of combinatorial CREs. The program uses Position Specific Scoring Matrices (PSSMs) representing possible transcription factor binding sites, in order to find synergistic combinations of CREs. The importance of combinatorial control is stated in a detailed review written by (Singh 1998). In the present study, PSSMs predicted by a motif finding program and gene expression data were used to predict combinatorial CREs. Similar bioinformatic approaches for combinatorial CREs prediction have already been reported. (Pilpel et al. 2001) assessed *Sacharomyces cerevisae* microarray data to find synergistic combinations of motifs that were statistically significant. Steps in the strategy followed by (Pilpel et al. 2001) for combinatorial CREs prediction are similar to the ones implemented in the present study. (Pilpel et al. 2001) first used a motif finding program to predict motifs that were further used to identify genes containing combinations of such motifs within their promoters. Expression of such genes was calculated using microarray data and finally statistically significant combinations were identified (Pilpel et al. 2001). One important difference between the present study and the one reported by (Pilpel et al. 2001) is that they did not include parameters such as motif orientation and order within promoters for combinatorial element prediction, which have been shown to have an effect on motif function (Werner 1999). Such spatial constraints were shown to have an effect on successful combinatorial element prediction also by (Beer and Tavazoie 2004), which indicates that they should be included when combinatorial elements are predicted.

For plants, bioinformatic identification of combinatorial elements has also been reported for *Arabidopsis thaliana*. (Cserháti et al. 2011) developed a statistical algorithm for the prediction of combinatorial motifs putatively responsive to abiotic stresses. The combinatorial elements were predicted by assessing if their presence in promoters of stress-induced genes is statistically higher than in the promoters of not-induced genes (Cserháti et al. 2011). That contrasts with the approach followed in the present study, where overall genomic occurrences are used for combinatorial element prediction, complemented by correlation with gene expression data, rather than only focusing on certain genes induced upon stresses. Although there are methodical

similarities between both studies, like a similar allowed wooble of ±5bp in the spacer length, the analysis presented here has several advantages over the analysis by (Cserháti et al. 2011). There were no constraints in how long the spacer length should be, as opposed to the maximum spacer of 52bp (Cserháti et al. 2011). PSSMs, instead of single sequences (Cserháti et al. 2011), were used to represent the CREs, which has been shown to accurately represent a TFBS. Also the length of the CREs forming the combinatorial elements was set to a fixed length of 5bp (Cserháti et al. 2011), whereas in this study the length of the motifs forming the elements was 5 to 10bp, which is the typical length of a CRE (Solovyev et al. 2010). Finally, not only abiotic stresses were taken into account to assess the putative functionality of the predicted CREs, but all 155 stresses from PathoPlant (including biotic stresses) were assessed.

Another study reporting the identification of combinatorial elements in *Arabidopsis thaliana* was carried out by (Zou et al. 2011). By analyzing 16 stresses (biotic- and abiotic-related), (Zou et al. 2011) developed prediction models that were used initially to predict 1215 single putatively functional CREs. It was demonstrated that by considering combinatorial control, the prediction models could be improved (Zou et al. 2011). However, spatial motif constraints such as spacers and relative orientation were not used for the prediction of combinatorial elements (Zou et al. 2011). These constraints have been shown to have an effect on combinatorial element functionality (Yu et al. 2006), and constitute an important difference between combinatorial elements from the present study and the ones reported by (Zou et al. 2011).

It has been shown that low affinity or weak TFBSs occur extensively in eukaryotic genomes, but interactions between such sites are largely not yet understood (Tanay 2006). Furthermore, (Gertz et al. 2009) reported functional combinations of low- and high-affinity binding sites interacting with each other in *Sacharomyces cerevisae*. Thus, a possible expansion of the program developed in this study would be to model weak interactions in the *Arabidopsis thaliana* genome. In this study, motifs predicted by the program MEME were used as input sequences for combinatorial element prediction. These motifs could represent high and low-affinity binding sites. By including experimental information about the affinity of the binding sites, which can be measured from high-throughput ChiP data (Tanay 2006), the presence of binding sites

combinations with high- and low-affinity could be assessed with the program developed in this study. In that way, the reported combinatorial element predictions could be further improved.

4.3.1 Characteristic spatial constraints

Combinatorial elements were predicted in this study with and without spatial constraints. Such constraints included spacer lengths, motif order and motif orientation. It was shown that predicting combinatorial elements without spatial constraints results in a very low number of predictions (see **Table 3.18** in page 89). This observation indicates that spacers seem to be very important for combinatorial element functionality, which correlates with observations made by (Yu et al. 2006). The majority of combinatorial elements predicted in the present study displayed characteristic short (<100bp) spacer lengths (see **Figure 3.22** in page 94). This correlates with observations made by (Yu et al. 2006), where it was demonstrated that 75% of combinatorial elements predicted displayed spacer distances shorter than 166bp. Nevertheless, long spacers were also observed for some of the combinatorial elements in the present study (see **Figure 3.21** in page 93). This was also found with the combinatorial elements predicted by (Yu et al. 2006), where 25% of the elements displayed a spacer length >166bp. As noted by (Yu et al. 2006), these longer distances could reflect interactions through DNA looping.

Among the combinatorial element with short spacer lengths (<100bp), certain element sets displayed characteristic spacer lengths that turned out to be even time point-specific. Elements putatively responsive to Flg22 1hr, have a clear and significant spacer length of 0-50bp and no characteristic length in the range of 51-100bp (see **Figure 3.19** in page 91). On the other hand, Flg22 4hr elements display a characteristic length of 51-100bp, contrasting with the results observed for Flg22 1hr (see **Figure 3.20** in page 92). Although it can be argued that the chosen range of 50bp to assess spacer lengths is somewhat arbitrary and that it can have an effect on the observed spatial patterns, it is nevertheless clear that the spacer lengths are very different for Flg22 1hr and 4hr responsive elements. This correlates with the observation made by (Cserháti et al. 2011) that biologically relevant combinatorial elements responsive to a

given stress display similar characteristic spacer lengths. In addition, also correlating with the observed results, (Yu et al. 2006) reported different characteristic distances for combinatorial elements under different conditions. Thus, the results imply that different arrangements of transcription factors could be specific for a given stress and even for a given time point, which could confer high specificity for a precise response to a stress. Experimental validation will be needed to prove that predicted combinatorial elements responsive to Flg22 1hr are not be responsive to Flg22 4hr and *vice versa*.

Another spatial constraint tested in this study was the order of motifs forming a combinatorial element. After testing several spatial constraints, (Beer and Tavazoie 2004) found in a case study that the motif order has a strong effect in the degree of correlation of genes containing the combinatorial element, implying that the motif order is important for combinatorial element prediction. In the present study it was demonstrated that, when compared with combinatorial elements having a specific motif order, elements without order increase the number of randomly expected combinatorial elements (see **Table 3.20** in page 97). In addition, although the number of predicted elements without order is higher, the variety of elements is not notably different in both sets. These observations together with the evidence of the order importance in other experiments (Beer and Tavazoie 2004) indicate that this spatial constraint is relevant for combinatorial element prediction.

Further spatial constraints measured were the orientations of the motifs to each other forming the combinatorial elements and their distance to the TSS. The importance of the motifs orientations was highlighted by (Yu et al. 2006). Although motifs forming certain combinatorial elements show specific relative orientation preferences, no clear general preference was observed for all elements predicted in this study (see **Table 3.19** in page 96). Thus, it can be observed that the orientation preferences are present for specific combinatorial elements instead of being a general characteristic. The importance of the distance to the TSS was pointed out by (Vardhanabhuti et al. 2007). In the present study, no general distance to the TSS could be observed for the whole set of predicted combinatorial elements (see **Figure 3.26** in page 100). However, this observation does not rule out the possibility that certain combinatorial elements may

have a characteristic distance to the TSS. It should then be tested if single combinatorial elements display these characteristic distances, and, should that be the case, the algorithm developed in the present study could be refined in order to identify combinatorial elements displaying similar distances to the TSS.

5 Summary

The goal of the present work was the development of bioinformatics methods for the identification of putatively functional stress-responsive *cis*-regulatory elements (CREs) in plants. For that purpose, microarray experiment data of the model plant *Arabidopsis thaliana* were used. A novel tool called *in silico* expression analysis was developed in the course of the present work. The tool correlates genome-wide promoter occurrences of a given sequence with microarray expression data stored in the PathoPlant database. It provides statistical values which serve to evaluate the probability of a sequence being responsive to a given stress. The newly developed tool was validated by analyzing known stress-responsive CREs. The Drought Responsive Element (DRE), pathogen-responsive CREs like the WRKY AGTTGACTAA, a Zn-deficiency responsive element and a salicylic acid responsive sequence all displayed significant p-values (<0.001) for the expected stresses. Also, as another validation approach, two sets of synthetic CREs from a high-throughput experimental approach were analyzed with the tool. One set with 3096 elements after treatment with Pep-25 (enriched set) and one untreated control set with 2801 elements. The analysis showed that the enriched set contained a higher proportion of fungal-responsive elements than the untreated control set.

The *in silico* expression analysis tool was used to identify putatively functional CREs within motif sets predicted by the motif-finding program MEME. 18 gene promoter sets corresponding to genes up-regulated upon: Chitooctaose, EF-Tu, Flg22, Pb-Oversupply, Zn-defficiency and Zn-Oversupply at different time points, were used as input sequences for MEME. The program yielded a total of 6700 motifs which were further used as input sequences for identification of CREs with the *in silico* expression analysis. This resulted in a set of 1014 putatively functional CREs. The novelty of the predicted CREs was assessed using the STAMP web server. The analysis revealed that among the CREs predicted to be responsive to biotic stresses, elements similar to known stress-responsive elements (WBOX and ABRE) were observed. In addition the abiotic elements displayed similarities to Zn-deficiency and Iron deficiency responsive elements. The redundancy among the predicted motifs was also assessed with STAMP, which classified the motifs into 261 different motif clusters. As an approach to reduce

the redundancy, i.e. increase the variety among predicted CREs, a new approach was developed. A set with all possible DNA 10mers (1,048,576 different sequences) was generated. The sequences within the set were used as input for the *in silico* expression analysis, which yielded information about the expression of genes containing each of those sequences within the promoters. The result of the *in silico* expression analysis was used to identify sequences putatively responsive to Flg22 and drought stresses. The sequences were ranked according to p-values and the top 30 sequences were further analyzed with STAMP. The sequences were clustered into 5 different groups, with the biggest group showing similarities to the pathogen-responsive element W-Box. Further similarities were observed with the CREs TELO-box, RBENTGA3, TGA1ANTPR1A and 3AF1BOXPSRBCS3. The Drought responsive elements showed similarities to the CREs ABRE3HVA1, SORLIP5AT, ABFs, ABF1, OCSGMHSP26A, AtMYB2, BOXLCOREDCPAL and WRECSAA01.

Two novel web tools were developed during the present study. One is an on-line version of the *in silico* expression analysis already publicly available. The tool allows the identification of genes containing a user-submitted sequence within promoters. Expression information of such genes, together with statistical analysis are given to the user, thus allowing the evaluation of possible sequence functionality. The tool is freely available at http://www.pathoplant.de/expression_analysis.php. Another web tool that will also be published is *cis*-elements. It allows a user to select a stress from the PathoPlant database in order to obtain putatively functional CREs associated to that stress. It is possible to define the specificity of the predicted CREs in order to filter the predicted CREs.

Finally a new program was developed for the prediction of combinatorial CREs. The program searched for motif combinations in the *Arabidopsis* gene promoters and calculated the expression of genes containing such combinations within the promoters. The statistical significance of the calculated expression was used to identify putatively functional combinatorial elements. The program identified 788 motif combinations with spatial constraints (spacer length, motif orientation and motif order) and only 12 combinations without constraints, indicating that such constraints are very important for motif prediction. A study of the spacer length frequency in the

set of combinatorial elements with spatial constraints revealed that the majority of elements displayed short spacer lengths (≤100 nucleotides).

The results show that the novel developed bioinformatics tools serve to predict CREs responsive to different biotic and abiotic stresses. Promising novel CRE candidates should be further experimentally analyzed.

6 References

Abe, H.; Yamaguchi-Shinozaki, K.; Urao, T.; Iwasaki, T.; Hosokawa, D.; Shinozaki, K. (1997): Role of arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. In: *Plant Cell* 9 (10), S. 1859–1868.

Aharoni, Asaph; Vorst, Oscar (2002): DNA microarrays for functional plant genomics. In: *Plant Mol. Biol.* 48 (1-2), S. 99–118.

Assunção, Ana G. L.; Herrero, Eva; Lin, Ya-Fen; Huettel, Bruno; Talukdar, Sangita; Smaczniak, Cezary et al. (2010): Arabidopsis thaliana transcription factors bZIP19 and bZIP23 regulate the adaptation to zinc deficiency. In: *Proc. Natl. Acad. Sci. U.S.A.* 107 (22), S. 10296–10301.

Bailey, T. L.; Elkan, C. (1994): Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proc Int Conf Intell Syst Mol Biol* 2, S. 28–36.

Bailey, T. L.; Elkan, C. (1995a): The value of prior knowledge in discovering motifs with MEME. In: *Proc Int Conf Intell Syst Mol Biol* 3, S. 21–29.

Bailey, Timothy L.; Elkan, Charles (1995b): Unsupervised learning of multiple motifs in biopolymers using expectation maximization. In: *Mach Learn* 21 (1-2), S. 51–80.

Bailey, Timothy L.; Boden, Mikael; Buske, Fabian A.; Frith, Martin; Grant, Charles E.; Clementi, Luca et al. (2009): MEME SUITE: tools for motif discovery and searching. In: *Nucleic Acids Res.* 37 (Web Server issue), S. W202-8.

Bailey, Timothy L.; Williams, Nadya; Misleh, Chris; Li, Wilfred W. (2006): MEME: discovering and analyzing DNA and protein sequence motifs. In: *Nucleic Acids Res.* 34 (Web Server issue), S. W369-73.

Bammler, Theodore; Beyer, Richard P.; Bhattacharya, Sanchita; Boorman, Gary A.; Boyles, Abee; Bradford, Blair U. et al. (2005): Standardizing global gene expression analysis between laboratories and across platforms. In: *Nat. Methods* 2 (5), S. 351–356.

Beer, Michael A.; Tavazoie, Saeed (2004): Predicting gene expression from sequence. In: *Cell* 117 (2), S. 185–198.

Blanco, Enrique; Messeguer, Xavier; Smith, Temple F.; Guigó, Roderic (2006): Transcription factor map alignment of promoter regions. In: *PLoS Comput. Biol.* 2 (5), S. e49.

Bostock, Richard M. (2005): Signal crosstalk and induced resistance: straddling the line between cost and benefit. In: *Annu Rev Phytopathol* 43, S. 545–580.

Bruce Alberts.; Alberts, Bruce (2008): Molecular biology of the cell. New York, [London: Garland Science; Taylor & Francis, distributor].

Bülow, Lorenz; Bolívar, Julio C.; Ruhe, Jonas; Brill, Yuri; Hehl, Reinhard (2012): 'MicroRNA Targets', a new AthaMap web-tool for genome-wide identification of miRNA targets in Arabidopsis thaliana. In: *BioData Min* 5 (1), S. 7.

Bülow, Lorenz; Brill, Yuri; Hehl, Reinhard (2010): AthaMap-assisted transcription factor target gene identification in Arabidopsis thaliana. In: *Database (Oxford)* 2010, S. baq034.

Bülow, Lorenz; Engelmann, Stefan; Schindler, Martin; Hehl, Reinhard (2009): AthaMap, integrating transcriptional and post-transcriptional data. In: *Nucleic Acids Res.* 37 (Database issue), S. D983-6.

Bülow, Lorenz; Schindler, Martin; Hehl, Reinhard (2007): PathoPlant: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. In: *Nucleic Acids Res.* 35 (Database issue), S. D841-5.

Bülow, Lorenz; Schindler, Martin; Choi, Claudia; Hehl, Reinhard (2004): PathoPlant: a database on plant-pathogen interactions. In: *In Silico Biol. (Gedrukt)* 4 (4), S. 529–536.

Busk, P. K.; Jensen, A. B.; Pagès, M. (1997): Regulatory elements in vivo in the promoter of the abscisic acid responsive gene rab17 from maize. In: *Plant J.* 11 (6), S. 1285–1295.

Bussemaker, H. J.; Li, H.; Siggia, E. D. (2001): Regulatory element detection using correlation with expression. In: *Nat. Genet.* 27 (2), S. 167–171.

Carey, M. (1998): The enhanceosome and transcriptional synergy. In: *Cell* 92 (1), S. 5–8.

Caselle, Michele; Di Cunto, Ferdinando; Provero, Paolo (2002): Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. In: *BMC Bioinformatics* 3, S. 7.

Chakravarthy, Suma; Tuori, Robert P.; D'Ascenzo, Mark D.; Fobert, Pierre R.; Despres, Charles; Martin, Gregory B. (2003): The tomato transcription factor Pti4 regulates defense-related gene expression via GCC box and non-GCC box cis elements. In: *Plant Cell* 15 (12), S. 3033–3050.

Chan, C. S.; Guo, L.; Shih, M. C. (2001): Promoter analysis of the nuclear gene encoding the chloroplast glyceraldehyde-3-phosphate dehydrogenase B subunit of Arabidopsis thaliana. In: *Plant Mol. Biol.* 46 (2), S. 131–141.

Chang, Wen-Chi; Lee, Tzong-Yi; Huang, Hsien-Da; Huang, His-Yuan; Pan, Rong-Long (2008): PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. In: *BMC Genomics* 9, S. 561.

Chinnusamy, Viswanathan; Schumaker, Karen; Zhu, Jian-Kang (2004): Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. In: *J. Exp. Bot.* 55 (395), S. 225–236.

Choi, H.; Hong, J.; Ha, J.; Kang, J.; Kim, S. Y. (2000): ABFs, a family of ABA-responsive element binding factors. In: *J. Biol. Chem.* 275 (3), S. 1723–1730.

Ciolkowski, Ingo; Wanke, Dierk; Birkenbihl, Rainer P.; Somssich, Imre E. (2008): Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. In: *Plant Mol. Biol.* 68 (1-2), S. 81–92.

Colcombet, Jean; Hirt, Heribert (2008): Arabidopsis MAPKs: a complex signalling network involved in multiple biological processes. In: *Biochem. J.* 413 (2), S. 217–226.

Cserháti, Mátyás; Turóczy, Zoltán; Zombori, Zoltán; Cserzo, Miklós; Dudits, Dénes; Pongor, Sándor; Györgyey, János (2011): Prediction of new abiotic stress genes in Arabidopsis thaliana and Oryza sativa according to enumeration-based statistical analysis. In: *Mol. Genet. Genomics* 285 (5), S. 375–391.

Cushman, J. C.; Bohnert, H. J. (2000): Genomic approaches to plant stress tolerance. In: *Curr. Opin. Plant Biol.* 3 (2), S. 117–124.

Das, Modan K.; Dai, Ho-Kwok (2007): A survey of DNA motif finding algorithms. In: *BMC Bioinformatics* 8 Suppl 7, S. S21.

Davuluri, Ramana V.; Sun, Hao; Palaniswamy, Saranyan K.; Matthews, Nicole; Molina, Carlos; Kurtz, Mike; Grotewold, Erich (2003): AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. In: *BMC Bioinformatics* 4, S. 25.

Debnath, Mousumi; Pandey, Mukeshwar; Bisen, P. S. (2011): An omics approach to understand the plant abiotic stress. In: *OMICS* 15 (11), S. 739–762.

Dong, Jixin; Chen, Chunhong; Chen, Zhixiang (2003): Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response. In: *Plant Mol. Biol.* 51 (1), S. 21–37.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. (1998): Cluster analysis and display of genome-wide expression patterns. In: *Proc. Natl. Acad. Sci. U.S.A.* 95 (25), S. 14863–14868.

Ellis, J. G.; Tokuhisa, J. G.; Llewellyn, D. J.; Bouchez, D.; Singh, K.; Dennis, E. S.; Peacock, W. J. (1993): Does the ocs-element occur as a functional component of the promoters of plant genes? In: *Plant J.* 4 (3), S. 433–443.

Eulgem, T.; Rushton, P. J.; Robatzek, S.; Somssich, I. E. (2000): The WRKY superfamily of plant transcription factors. In: *Trends Plant Sci.* 5 (5), S. 199–206.

Eulgem, T.; Rushton, P. J.; Schmelzer, E.; Hahlbrock, K.; Somssich, I. E. (1999): Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. In: *EMBO J.* 18 (17), S. 4689–4699.

Eulgem, Thomas; Somssich, Imre E. (2007): Networks of WRKY transcription factors in defense signaling. In: *Curr. Opin. Plant Biol.* 10 (4), S. 366–371.

Fan, Jun; Hill, Lionel; Crooks, Casey; Doerner, Peter; Lamb, Chris (2009): Abscisic acid has a key role in modulating diverse plant-pathogen interactions. In: *Plant Physiol.* 150 (4), S. 1750–1761.

Firestein, Gary S.; Pisetsky, David S. (2002): DNA microarrays: boundless technology or bound by technology? Guidelines for studies using microarray technology. In: *Arthritis Rheum.* 46 (4), S. 859–861.

Fujita, Miki; Fujita, Yasunari; Noutoshi, Yoshiteru; Takahashi, Fuminori; Narusaka, Yoshihiro; Yamaguchi-Shinozaki, Kazuko; Shinozaki, Kazuo (2006): Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. In: *Curr. Opin. Plant Biol.* 9 (4), S. 436–442.

Fujita, Yasunari; Fujita, Miki; Satoh, Rie; Maruyama, Kyonoshin; Parvez, Mohammad M.; Seki, Motoaki et al. (2005): AREB1 is a transcription activator of novel ABRE-dependent ABA signaling that enhances drought stress tolerance in Arabidopsis. In: *Plant Cell* 17 (12), S. 3470–3488.

Fukazawa, J.; Sakai, T.; Ishida, S.; Yamaguchi, I.; Kamiya, Y.; Takahashi, Y. (2000): Repression of shoot growth, a bZIP transcriptional activator, regulates cell elongation by controlling the level of gibberellins. In: *Plant Cell* 12 (6), S. 901–915.

Galuschka, Claudia; Schindler, Martin; Bülow, Lorenz; Hehl, Reinhard (2007): AthaMap web tools for the analysis and identification of co-regulated genes. In: *Nucleic Acids Res.* 35 (Database issue), S. D857-62.

Genoud; Métraux (1999): Crosstalk in plant cell signaling: structure and function of the genetic network. In: *Trends Plant Sci.* 4 (12), S. 503–507.

Gertz, Jason; Siggia, Eric D.; Cohen, Barak A. (2009): Analysis of combinatorial cis-regulation in synthetic and genomic promoters. In: *Nature* 457 (7226), S. 215–218.

Girke, T.; Todd, J.; Ruuska, S.; White, J.; Benning, C.; Ohlrogge, J. (2000): Microarray analysis of developing Arabidopsis seeds. In: *Plant Physiol.* 124 (4), S. 1570–1581.

Green, Michael R. (2005): Eukaryotic transcription activation: right on target. In: *Mol. Cell* 18 (4), S. 399–402.

Group M. (2002): Mitogen-activated protein kinase cascades in plants: a new nomenclature. In: *Trends Plant Sci.* 7 (7), S. 301–308.

Guo, An-Yuan; Chen, Xin; Gao, Ge; Zhang, He; Zhu, Qi-Hui; Liu, Xiao-Chuan et al. (2008): PlantTFDB: a comprehensive plant transcription factor database. In: *Nucleic Acids Res.* 36 (Database issue), S. D966-9.

He, Xin; Ling, Xu; Sinha, Saurabh (2009): Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. In: *PLoS Comput. Biol.* 5 (3), S. e1000299.

Hehl, Reinhard; Bülow, Lorenz (2008): Internet Resources for Gene Expression Analysis in Arabidopsis thaliana. In: *Curr. Genomics* 9 (6), S. 375–380.

Higo, K.; Ugawa, Y.; Iwamoto, M.; Korenaga, T. (1999): Plant cis-acting regulatory DNA elements (PLACE) database: 1999. In: *Nucleic Acids Res.* 27 (1), S. 297–300.

Hu, Jianjun; Li, Bin; Kihara, Daisuke (2005): Limitations and potentials of current motif discovery algorithms. In: *Nucleic Acids Res.* 33 (15), S. 4899–4913.

Hudson, Matthew E.; Quail, Peter H. (2003): Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. In: *Plant Physiol.* 133 (4), S. 1605–1616.

Hughes, J. D.; Estep, P. W.; Tavazoie, S.; Church, G. M. (2000): Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. In: *J. Mol. Biol.* 296 (5), S. 1205–1214.

Irizarry, Rafael A.; Bolstad, Benjamin M.; Collin, Francois; Cope, Leslie M.; Hobbs, Bridget; Speed, Terence P. (2003): Summaries of Affymetrix GeneChip probe level data. In: *Nucleic Acids Res.* 31 (4), S. e15.

Janaki, Chintalapati; Joshi, Rajendra R. (2004): Motif detection in Arabidopsis: correlation with gene expression data. In: *In Silico Biol. (Gedrukt)* 4 (2), S. 149–161.

Johnson, Christopher; Boden, Erin; Arias, Jonathan (2003): Salicylic acid and NPR1 induce the recruitment of trans-activating TGA factors to a defense gene promoter in Arabidopsis. In: *Plant Cell* 15 (8), S. 1846–1858.

Kel, O. V.; Romaschenko, A. G.; Kel, A. E.; Wingender, E.; Kolchanov, N. A. (1995): A compilation of composite regulatory elements affecting gene transcription in vertebrates. In: *Nucleic Acids Res.* 23 (20), S. 4097–4103.

Kim, Hyoun-Joung; Kim, Yun-Kyoung; Park, Jin-Young; Kim, Jungmook (2002): Light signalling mediated by phytochrome plays an important role in cold-induced gene expression through the C-repeat/dehydration responsive element (C/DRE) in Arabidopsis thaliana. In: *Plant J.* 29 (6), S. 693–704.

Kim, June-Sik; Mizoi, Junya; Yoshida, Takuya; Fujita, Yasunari; Nakajima, Jun; Ohori, Teppei et al. (2011): An ABRE promoter sequence is involved in osmotic stress-responsive expression of the DREB2A gene, which encodes a transcription factor regulating drought-inducible genes in Arabidopsis. In: *Plant Cell Physiol.* 52 (12), S. 2136–2146.

Kobayashi, Takanori; Nakayama, Yuko; Itai, Reiko Nakanishi; Nakanishi, Hiromi; Yoshihara, Toshihiro; Mori, Satoshi; Nishizawa, Naoko K. (2003): Identification of novel cis-acting elements, IDE1 and IDE2, of the barley IDS2 gene promoter conferring iron-deficiency-inducible, root-specific expression in heterogeneous tobacco plants. In: *Plant J.* 36 (6), S. 780–793.

Lam, E.; Kano-Murakami, Y.; Gilmartin, P.; Niner, B.; Chua, N. H. (1990): A metal-dependent DNA-binding protein interacts with a constitutive element of a light-responsive promoter. In: *Plant Cell* 2 (9), S. 857–866.

Langenau, Frank (2001): Microsoft SQL Server 2000. Für Datenbankadministration and -entwicklung. München/Germany: Markt-und-Technik-Verl.

Lashkari, D. A.; DeRisi, J. L.; McCusker, J. H.; Namath, A. F.; Gentile, C.; Hwang, S. Y. et al. (1997): Yeast microarrays for genome wide parallel genetic and gene expression analysis. In: *Proc. Natl. Acad. Sci. U.S.A.* 94 (24), S. 13057–13062.

Lee, T. I.; Young, R. A. (2000): Transcription of eukaryotic protein-coding genes. In: *Annu. Rev. Genet.* 34, S. 77–137.

Levine, Michael; Tjian, Robert (2003): Transcription regulation and animal diversity. In: *Nature* 424 (6945), S. 147–151.

Lipshutz, R. J.; Fodor, S. P.; Gingeras, T. R.; Lockhart, D. J. (1999): High density synthetic oligonucleotide arrays. In: *Nat. Genet.* 21 (1 Suppl), S. 20–24.

Liu, X. Shirley; Brutlag, Douglas L.; Liu, Jun S. (2002): An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. In: *Nat. Biotechnol.* 20 (8), S. 835–839.

Liu, X.; Brutlag, D. L.; Liu, J. S. (2001): BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In: *Pac Symp Biocomput*, S. 127–138.

Lockhart, D. J.; Winzeler, E. A. (2000): Genomics, gene expression and DNA arrays. In: *Nature* 405 (6788), S. 827–836.

Ma, Shisong; Gong, Qingqiu; Bohnert, Hans J. (2006): Dissecting salt stress pathways. In: *J. Exp. Bot.* 57 (5), S. 1097–1107.

Maeda, Kazuhiro; Kimura, Soichi; Demura, Taku; Takeda, Junko; Ozeki, Yoshihiro (2005): DcMYB1 acts as a transcriptional activator of the carrot phenylalanine ammonia-lyase gene (DcPAL1) in response to elicitor treatment, UV-B irradiation and the dilution effect. In: *Plant Mol. Biol.* 59 (5), S. 739–752.

Mahony, Shaun; Benos, Panayiotis V. (2007): STAMP: a web tool for exploring DNA-binding motif similarities. In: *Nucleic Acids Res.* 35 (Web Server issue), S. W253-8.

Mahony, Shaun; Auron, Philip E.; Benos, Panayiotis V. (2007): DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. In: *PLoS Comput. Biol.* 3 (3), S. e61.

Maruyama, Kyonoshin; Sakuma, Yoh; Kasuga, Mie; Ito, Yusuke; Seki, Motoaki; Goda, Hideki et al. (2004): Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. In: *Plant J.* 38 (6), S. 982–993.

Maruyama-Nakashita, Akiko; Nakamura, Yumiko; Watanabe-Takahashi, Akiko; Inoue, Eri; Yamaya, Tomoyuki; Takahashi, Hideki (2005): Identification of a novel cis-acting element conferring sulfur deficiency response in Arabidopsis roots. In: *Plant J.* 42 (3), S. 305–314.

Mauch-Mani, Brigitte; Mauch, Felix (2005): The role of abscisic acid in plant-pathogen interactions. In: *Curr. Opin. Plant Biol.* 8 (4), S. 409–414.

Mészáros, Tamás; Helfer, Anne; Hatzimasoura, Elizabeth; Magyar, Zoltán; Serazetdinova, Liliya; Rios, Gabino et al. (2006): The Arabidopsis MAP kinase kinase MKK1 participates in defence responses to the bacterial elicitor flagellin. In: *Plant J.* 48 (4), S. 485–498.

Mohanty, Bijayalaxmi; Krishnan, S. P. T.; Swarup, Sanjay; Bajic, Vladimir B. (2005): Detection and preliminary analysis of motifs in promoters of anaerobically induced genes of different plant species. In: *Ann. Bot.* 96 (4), S. 669–681.

Mundy, J.; Yamaguchi-Shinozaki, K.; Chua, N. H. (1990): Nuclear proteins bind conserved elements in the abscisic acid-responsive promoter of a rice rab gene. In: *Proc. Natl. Acad. Sci. U.S.A.* 87 (4), S. 1406–1410.

Nagao, R. T.; Goekjian, V. H.; Hong, J. C.; Key, J. L. (1993): Identification of protein-binding DNA sequences in an auxin-regulated gene of soybean. In: *Plant Mol. Biol.* 21 (6), S. 1147–1162.

Nakashima, Kazuo; Ito, Yusuke; Yamaguchi-Shinozaki, Kazuko (2009): Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. In: *Plant Physiol.* 149 (1), S. 88–95.

Narusaka, Yoshihiro; Narusaka, Mari; Seki, Motoaki; Umezawa, Taishi; Ishida, Junko; Nakajima, Maiko et al. (2004): Crosstalk in the responses to abiotic and biotic stresses in Arabidopsis: analysis of gene expression in cytochrome P450 gene superfamily by cDNA microarray. In: *Plant Mol. Biol.* 55 (3), S. 327–342.

Needleman, S. B.; Wunsch, C. D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. In: *J. Mol. Biol.* 48 (3), S. 443–453.

Ono, A.; Izawa, T.; Chua, N. H.; Shimamoto, K. (1996): The rab16B promoter of rice contains two distinct abscisic acid-responsive elements. In: *Plant Physiol.* 112 (2), S. 483–491.

Pabo, C. O.; Sauer, R. T. (1992): Transcription factors: structural families and principles of DNA recognition. In: *Annu. Rev. Biochem.* 61, S. 1053–1095.

Palm, C. J.; Costa, M. A.; An, G.; Ryan, C. A. (1990): Wound-inducible nuclear protein binds DNA fragments that regulate a proteinase inhibitor II gene from potato. In: *Proc. Natl. Acad. Sci. U.S.A.* 87 (2), S. 603–607.

Pandey, Shree P.; Somssich, Imre E. (2009): The role of WRKY transcription factors in plant immunity. In: *Plant Physiol.* 150 (4), S. 1648–1655.

Pavesi, Giulio; Mereghetti, Paolo; Mauri, Giancarlo; Pesole, Graziano (2004): Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. In: *Nucleic Acids Res.* 32 (Web Server issue), S. W199-203.

Pierstorff, Nora; Bergman, Casey M.; Wiehe, Thomas (2006): Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. In: *Bioinformatics* 22 (23), S. 2858–2864.

Pilpel, Y.; Sudarsanam, P.; Church, G. M. (2001): Identifying regulatory networks by combinatorial analysis of promoter elements. In: *Nat. Genet.* 29 (2), S. 153–159.

Priest, Henry D.; Filichkin, Sergei A.; Mockler, Todd C. (2009): Cis-regulatory elements in plant cell signaling. In: *Curr. Opin. Plant Biol.* 12 (5), S. 643–649.

Qin, Feng; Shinozaki, Kazuo; Yamaguchi-Shinozaki, Kazuko (2011): Achievements and challenges in understanding plant abiotic stress responses and tolerance. In: *Plant Cell Physiol.* 52 (9), S. 1569–1582.

Reis-Filho, Jorge S. (2009): Next-generation sequencing. In: *Breast Cancer Res.* 11 Suppl 3, S. S12.

Reményi, Attila; Schöler, Hans R.; Wilmanns, Matthias (2004): Combinatorial control of gene expression. In: *Nat. Struct. Mol. Biol.* 11 (9), S. 812–815.

Rodriguez, Maria Cristina Suarez; Petersen, Morten; Mundy, John (2010): Mitogen-activated protein kinase signaling in plants. In: *Annu Rev Plant Biol* 61, S. 621–649.

Roeder, R. G. (1996): The role of general initiation factors in transcription by RNA polymerase II. In: *Trends Biochem. Sci.* 21 (9), S. 327–335.

Rushton, Paul J.; Somssich, Imre E.; Ringler, Patricia; Shen, Qingxi J. (2010): WRKY transcription factors. In: *Trends Plant Sci.* 15 (5), S. 247–258.

Sandve, Geir Kjetil; Drabløs, Finn (2006): A survey of motif discovery methods in an integrated framework. In: *Biol. Direct* 1, S. 11.

Schaffer, R.; Landgraf, J.; Accerbi, M.; Simon, V.; Larson, M.; Wisman, E. (2001): Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. In: *Plant Cell* 13 (1), S. 113–123.

Schenk, P. M.; Kazan, K.; Wilson, I.; Anderson, J. P.; Richmond, T.; Somerville, S. C.; Manners, J. M. (2000): Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. In: *Proc. Natl. Acad. Sci. U.S.A.* 97 (21), S. 11655–11660.

Schildt, Herbert (2011): Java. The complete reference. 8. Aufl. New York: McGraw-Hill.

Schleif, R. (1992): DNA looping. In: *Annu. Rev. Biochem.* 61, S. 199–223.

Seki, Motoaki; Narusaka, Mari; Ishida, Junko; Nanjo, Tokihiko; Fujita, Miki; Oono, Youko et al. (2002): Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. In: *Plant J.* 31 (3), S. 279–292.

Shen, Q.; Zhang, P.; Ho, T. H.D. (1996): Modular Nature of Abscisic Acid (ABA) Response Complexes: Composite Promoter Units That Are Necessary and Sufficient for ABA Induction of Gene Expression in Barley. In: *The Plant Cell Online* 8 (7), S. 1107–1119.

Shendure, Jay (2008): The beginning of the end for microarrays? In: *Nat. Methods* 5 (7), S. 585–587.

Shinozaki, K.; Yamaguchi-Shinozaki, K. (2000): Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. In: *Curr. Opin. Plant Biol.* 3 (3), S. 217–223.

Shinwari, Z. K.; Nakashima, K.; Miura, S.; Kasuga, M.; Seki, M.; Yamaguchi-Shinozaki, K.; Shinozaki, K. (1998): An Arabidopsis gene family encoding DRE/CRT binding proteins involved in low-temperature-responsive gene expression. In: *Biochem. Biophys. Res. Commun.* 250 (1), S. 161–170.

Siddharthan, Rahul; Siggia, Eric D.; van Nimwegen, Erik (2005): PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. In: *PLoS Comput. Biol.* 1 (7), S. e67.

Singh, K. B. (1998): Transcriptional regulation in plants: the importance of combinatorial control. In: *Plant Physiol.* 118 (4), S. 1111–1120.

Solano, R.; Nieto, C.; Avila, J.; Cañas, L.; Diaz, I.; Paz-Ares, J. (1995): Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from Petunia hybrida. In: *EMBO J.* 14 (8), S. 1773–1784.

Solovyev, Victor V.; Shahmuradov, Ilham A.; Salamov, Asaf A. (2010): Identification of promoter regions and regulatory sites. In: *Methods Mol. Biol.* 674, S. 57–83.

Steffens, Nils Ole; Galuschka, Claudia; Schindler, Martin; Bülow, Lorenz; Hehl, Reinhard (2004): AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. In: *Nucleic Acids Res.* 32 (Database issue), S. D368-72.

Steffens, Nils Ole; Galuschka, Claudia; Schindler, Martin; Bülow, Lorenz; Hehl, Reinhard (2005): AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in Arabidopsis thaliana. In: *Nucleic Acids Res.* 33 (Web Server issue), S. W397-402.

Stockinger, E. J.; Gilmour, S. J.; Thomashow, M. F. (1997): Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. In: *Proc. Natl. Acad. Sci. U.S.A.* 94 (3), S. 1035–1040.

Strompen, G.; Grüner, R.; Pfitzner, U. M. (1998): An as-1-like motif controls the level of expression of the gene for the pathogenesis-related protein 1a from tobacco. In: *Plant Mol. Biol.* 37 (5), S. 871–883.

Sutoh, Keita; Yamauchi, Daisuke (2003): Two cis-acting elements necessary and sufficient for gibberellin-upregulated proteinase expression in rice seeds. In: *Plant J.* 34 (5), S. 635–645.

Swarbreck, David; Wilks, Christopher; Lamesch, Philippe; Berardini, Tanya Z.; Garcia-Hernandez, Margarita; Foerster, Hartmut et al. (2008): The Arabidopsis Information Resource (TAIR): gene structure and function annotation. In: *Nucleic Acids Res.* 36 (Database issue), S. D1009-14.

Swindell, William R. (2006): The association among gene expression responses to nine abiotic stress treatments in Arabidopsis thaliana. In: *Genetics* 174 (4), S. 1811–1824.

Tagle, D. A.; Koop, B. F.; Goodman, M.; Slightom, J. L.; Hess, D. L.; Jones, R. T. (1988): Embryonic epsilon and gamma globin genes of a prosimian primate (Galago

crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. In: *J. Mol. Biol.* 203 (2), S. 439–455.

Tamura, Koichiro; Peterson, Daniel; Peterson, Nicholas; Stecher, Glen; Nei, Masatoshi; Kumar, Sudhir (2011): MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. In: *Mol. Biol. Evol.* 28 (10), S. 2731–2739.

Tanay, Amos (2006): Extensive low-affinity transcriptional interactions in the yeast genome. In: *Genome Res.* 16 (8), S. 962–972.

Thijs, Gert; Marchal, Kathleen; Lescot, Magali; Rombauts, Stephane; Moor, Bart de; Rouzé, Pierre; Moreau, Yves (2002): A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. In: *J. Comput. Biol.* 9 (2), S. 447–464.

Tompa, Martin; Li, Nan; Bailey, Timothy L.; Church, George M.; Moor, Bart de; Eskin, Eleazar et al. (2005): Assessing computational tools for the discovery of transcription factor binding sites. In: *Nat. Biotechnol.* 23 (1), S. 137–144.

Tremousaygue, D.; Manevski, A.; Bardet, C.; Lescure, N.; Lescure, B. (1999): Plant interstitial telomere motifs participate in the control of gene expression in root meristems. In: *Plant J.* 20 (5), S. 553–561.

Ulker, Bekir; Somssich, Imre E. (2004): WRKY transcription factors: from DNA binding towards biological function. In: *Curr. Opin. Plant Biol.* 7 (5), S. 491–498.

Ulmasov, T.; Liu, Z. B.; Hagen, G.; Guilfoyle, T. J. (1995): Composite structure of auxin response elements. In: *Plant Cell* 7 (10), S. 1611–1623.

Uno, Y.; Furihata, T.; Abe, H.; Yoshida, R.; Shinozaki, K.; Yamaguchi-Shinozaki, K. (2000): Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. In: *Proc. Natl. Acad. Sci. U.S.A.* 97 (21), S. 11632–11637.

van Hal, N. L.; Vorst, O.; van Houwelingen, A. M.; Kok, E. J.; Peijnenburg, A.; Aharoni, A. et al. (2000): The application of DNA microarrays in gene expression analysis. In: *J. Biotechnol.* 78 (3), S. 271–280.

Vardhanabhuti, Saran; Wang, Junwen; Hannenhalli, Sridhar (2007): Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. In: *Nucleic Acids Res.* 35 (10), S. 3203–3213.

Wang, J.; Ellwood, K.; Lehman, A.; Carey, M. F.; She, Z. S. (1999): A mathematical model for synergistic eukaryotic gene activation. In: *J. Mol. Biol.* 286 (2), S. 315–325.

Werner, T. (1999): Models for prediction and recognition of eukaryotic promoters. In: *Mamm. Genome* 10 (2), S. 168–175.

Xue, Gang-Ping (2003): The DNA-binding activity of an AP2 transcriptional activator HvCBF2 involved in regulation of low-temperature responsive genes in barley is modulated by temperature. In: *Plant J.* 33 (2), S. 373–383.

Yamaguchi-Shinozaki, K.; Shinozaki, K. (1994): A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. In: *Plant Cell* 6 (2), S. 251–264.

Yilmaz, Alper; Mejia-Guerra, Maria Katherine; Kurz, Kyle; Liang, Xiaoyu; Welch, Lonnie; Grotewold, Erich (2011): AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. In: *Nucleic Acids Res.* 39 (Database issue), S. D1118-22.

Yoshida, Takuya; Fujita, Yasunari; Sayama, Hiroko; Kidokoro, Satoshi; Maruyama, Kyonoshin; Mizoi, Junya et al. (2010): AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. In: *Plant J.* 61 (4), S. 672–685.

Yu, D.; Chen, C.; Chen, Z. (2001): Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression. In: *Plant Cell* 13 (7), S. 1527–1540.

Yu, Xueping; Lin, Jimmy; Masuda, Tomohiro; Esumi, Noriko; Zack, Donald J.; Qian, Jiang (2006): Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. In: *Nucleic Acids Res.* 34 (3), S. 917–927.

Yuh, C. H.; Bolouri, H.; Davidson, E. H. (1998): Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. In: *Science* 279 (5358), S. 1896–1902.

Zipfel, Cyril; Kunze, Gernot; Chinchilla, Delphine; Caniard, Anne; Jones, Jonathan D. G.; Boller, Thomas; Felix, Georg (2006): Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. In: *Cell* 125 (4), S. 749–760.

Zou, Cheng; Sun, Kelian; Mackaluso, Joshua D.; Seddon, Alexander E.; Jin, Rong; Thomashow, Michael F.; Shiu, Shin-Han (2011): Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. In: *Proc. Natl. Acad. Sci. U.S.A.* 108 (36), S. 14992–14997.

7 Appendix

7.1 Complete list of Microarray experiments implemented in PathoPlant

Table 7.1: Microarray experiments stored in the PathoPlant database

Stress	Class	Array type	No. of records	No. of genes	Expression set
Al-oversupplied roots	Abiotic stress	AFGC	18703	6880	1005823537
Cold-stressed roots 0.5hr	Abiotic stress	Affy ATH1	29790	15192	138
Cold-stressed roots 12hr	Abiotic stress	Affy ATH1	28810	14759	138
Cold-stressed roots 1hr	Abiotic stress	Affy ATH1	29801	15145	138
Cold-stressed roots 24hr	Abiotic stress	Affy ATH1	27878	14372	138
Cold-stressed roots 3hr	Abiotic stress	Affy ATH1	29719	15274	138
Cold-stressed roots 6hr	Abiotic stress	Affy ATH1	29635	15153	138
Cold-stressed shoots 0.5hr	Abiotic stress	Affy ATH1	25863	13291	138
Cold-stressed shoots 12hr	Abiotic stress	Affy ATH1	24808	12753	138
Cold-stressed shoots 1hr	Abiotic stress	Affy ATH1	25219	13169	138
Cold-stressed shoots 24hr	Abiotic stress	Affy ATH1	24296	12569	138
Cold-stressed shoots 3hr	Abiotic stress	Affy ATH1	26522	13628	138
Cold-stressed shoots 6hr	Abiotic stress	Affy ATH1	25470	13114	138
Drought-stressed roots 0.25hr	Abiotic stress	Affy ATH1	30197	15362	141
Drought-stressed roots 0.5hr	Abiotic stress	Affy ATH1	29937	15257	141
Drought-stressed roots 12hr	Abiotic stress	Affy ATH1	30310	15392	141
Drought-stressed roots 1hr	Abiotic stress	Affy ATH1	29874	15209	141
Drought-stressed roots 24hr	Abiotic stress	Affy ATH1	30316	15417	141
Drought-stressed roots 3hr	Abiotic stress	Affy ATH1	30104	15291	141
Drought-stressed roots 6hr	Abiotic stress	Affy ATH1	29820	15218	141
Drought-stressed shoots 0.25hr	Abiotic stress	Affy ATH1	25859	13309	141
Drought-stressed shoots 0.5hr	Abiotic stress	Affy ATH1	26096	13423	141

Drought-stressed shoots 12hr	Abiotic stress	Affy ATH1	27182	13878	141
Drought-stressed shoots 1hr	Abiotic stress	Affy ATH1	26519	13678	141
Drought-stressed shoots 24hr	Abiotic stress	Affy ATH1	25744	13191	141
Drought-stressed shoots 3hr	Abiotic stress	Affy ATH1	26954	13812	141
Drought-stressed shoots 6hr	Abiotic stress	Affy ATH1	26042	13373	141
Osmotic-stressed roots 0.5hr	Abiotic stress	Affy ATH1	29587	15112	139
Osmotic-stressed roots 12hr	Abiotic stress	Affy ATH1	28474	14619	139
Osmotic-stressed roots 1hr	Abiotic stress	Affy ATH1	29411	15008	139
Osmotic-stressed roots 24hr	Abiotic stress	Affy ATH1	29105	15004	139
Osmotic-stressed roots 3hr	Abiotic stress	Affy ATH1	29436	15054	139
Osmotic-stressed roots 6hr	Abiotic stress	Affy ATH1	28848	14897	139
Osmotic-stressed shoots 0.5hr	Abiotic stress	Affy ATH1	25687	13245	139
Osmotic-stressed shoots 12hr	Abiotic stress	Affy ATH1	25708	13193	139
Osmotic-stressed shoots 1hr	Abiotic stress	Affy ATH1	26419	13603	139
Osmotic-stressed shoots 24hr	Abiotic stress	Affy ATH1	24738	12787	139
Osmotic-stressed shoots 3hr	Abiotic stress	Affy ATH1	26173	13519	139
Osmotic-stressed shoots 6hr	Abiotic stress	Affy ATH1	25217	13012	139
Pb-oversupplied (25ppm) leaves	Abiotic stress	Affy 8K	4160	3865	19
Pb-oversupplied (25ppm) roots	Abiotic stress	Affy 8K	4668	4352	19
Pb-oversupplied (50ppm) leaves	Abiotic stress	Affy 8K	4388	4085	19
Pb-oversupplied (50ppm) roots	Abiotic stress	Affy 8K	4600	4289	19
Salt-stressed roots 0.5hr	Abiotic stress	Affy ATH1	29889	15217	140
Salt-stressed roots 12hr	Abiotic stress	Affy ATH1	28800	15035	140
Salt-stressed roots 1hr	Abiotic stress	Affy ATH1	29449	15014	140
Salt-stressed roots 24hr	Abiotic stress	Affy ATH1	29630	15180	140
Salt-stressed roots 3hr	Abiotic stress	Affy ATH1	29586	15088	140
Salt-stressed roots 6hr	Abiotic stress	Affy ATH1	28351	14606	140
Salt-stressed shoots 0.5hr	Abiotic stress	Affy ATH1	25917	13268	140
Salt-stressed shoots 12hr	Abiotic stress	Affy ATH1	26640	13683	140
Salt-stressed shoots 1hr	Abiotic stress	Affy	25649	13163	140

		ATH1			
Salt-stressed shoots 24hr	Abiotic stress	Affy ATH1	25820	13254	140
Salt-stressed shoots 3hr	Abiotic stress	Affy ATH1	25919	13373	140
Salt-stressed shoots 6hr	Abiotic stress	Affy ATH1	25851	13313	140
Zn-deficient roots	Abiotic stress	Affy ATH1	27831	14241	
Zn-deficient roots A. halleri	Abiotic stress	Affy ATH1	10087	10037	
Zn-deficient roots vs. resupplied Zn	Abiotic stress	Affy ATH1	27552	14221	
Zn-deficient shoots	Abiotic stress	Affy ATH1	23774	12562	
Zn-deficient shoots A. halleri	Abiotic stress	Affy ATH1	8755	8720	
Zn-deficient shoots vs. resupplied Zn	Abiotic stress	Affy ATH1	25204	13432	
Zn-oversupplied roots 2hr	Abiotic stress	Affy ATH1	25988	13287	
Zn-oversupplied roots 8hr	Abiotic stress	Affy ATH1	25668	13119	
Zn-oversupplied shoots 8hr	Abiotic stress	Affy ATH1	20832	11015	
Zn-resupplied roots 2hr vs. deficient Zn	Abiotic stress	Affy ATH1	27552	14221	
Zn-resupplied roots 2hr vs. suffficient Zn	Abiotic stress	Affy ATH1	27401	14069	
Zn-resupplied shoots 8hr vs. deficient Zn	Abiotic stress	Affy ATH1	25204	13432	
Zn-resupplied shoots 8hr vs. suffficient Zn	Abiotic stress	Affy ATH1	24165	12674	
P. syringae pv. maculicola 16hpi	Bacterial pathogen	Affy ATH1	23806	12438	1008031517
P. syringae pv. maculicola 24hpi	Bacterial pathogen	Affy ATH1	23964	12411	1008031517
P. syringae pv. maculicola 48hpi	Bacterial pathogen	Affy ATH1	24057	12634	1008031517
P. syringae pv. maculicola 4hpi	Bacterial pathogen	Affy ATH1	23661	12461	1008031517
P. syringae pv. maculicola 8hpi	Bacterial pathogen	Affy ATH1	23967	12534	1008031517
P. syringae pv. maculicola avrRpt2-16hpi	Bacterial pathogen	Affy ATH1	24144	12639	1008031517
P. syringae pv. maculicola avrRpt2-24hpi	Bacterial pathogen	Affy ATH1	24020	12462	1008031517
P. syringae pv. maculicola avrRpt2-48hpi	Bacterial pathogen	Affy ATH1	24172	12605	1008031517
P. syringae pv. maculicola avrRpt2-4hpi	Bacterial pathogen	Affy ATH1	23819	12529	1008031517
P. syringae pv. maculicola avrRpt2-8hpi	Bacterial pathogen	Affy ATH1	23790	12490	1008031517
P. syringae pv. phaseolicola 24hpi	Bacterial pathogen	Affy ATH1	34281	12129	1007966202
P. syringae pv. phaseolicola 2hpi	Bacterial	Affy	34364	12214	10079662

	pathogen	ATH1			02
P. syringae pv. phaseolicola 6hpi	Bacterial pathogen	Affy ATH1	36529	12773	10079662 02
P. syringae pv. tomato 24hpi	Bacterial pathogen	Affy ATH1	31636	11421	10079662 02
P. syringae pv. tomato 2hpi	Bacterial pathogen	Affy ATH1	34559	12209	10079662 02
P. syringae pv. tomato 6hpi	Bacterial pathogen	Affy ATH1	37126	13042	10079662 02
P. syringae pv. tomato avrRpm1 24hpi	Bacterial pathogen	Affy ATH1	33762	12021	10079662 02
P. syringae pv. tomato avrRpm1 2hpi	Bacterial pathogen	Affy ATH1	34701	12277	10079662 02
P. syringae pv. tomato avrRpm1 6hpi	Bacterial pathogen	Affy ATH1	36160	12802	10079662 02
P. syringae pv. tomato hrcC- 24hpi	Bacterial pathogen	Affy ATH1	34443	12154	10079662 02
P. syringae pv. tomato hrcC- 2hpi	Bacterial pathogen	Affy ATH1	34530	12118	10079662 02
P. syringae pv. tomato hrcC- 6hpi	Bacterial pathogen	Affy ATH1	37037	12880	10079662 02
X. campestris	Bacterial pathogen	AFGC	17911	6872	10058235 36
Inflorescence vs. shoot apex, vegetative	Development	Affy ATH1	40483	14000	153
Inflorescence vs. young leaves	Development	Affy ATH1	39889	13852	153
Chitin 10min	Elicitor	Carnegie	3949	1590	10058236 05
Chitin 1hr	Elicitor	Carnegie	3847	1568	10058236 05
Chitin 24hr	Elicitor	Carnegie	3665	1527	10058236 05
Chitin 30min	Elicitor	Carnegie	3915	1583	10058236 05
Chitin 3hr	Elicitor	Carnegie	3886	1574	10058236 05
Chitin 6hr	Elicitor	Carnegie	3895	1591	10058236 05
Chitooctaose	Elicitor	Affy ATH1	41687	14502	GSE8319
EF-Tu 30min	Elicitor	Affy ATH1	28792	14685	E-MEXP-547
EF-Tu 60min	Elicitor	Affy ATH1	28369	14511	E-MEXP-547
Flg22 (P. syringae) 1hr	Elicitor	Affy ATH1	31765	11544	10080807 27
Flg22 (P. syringae) 4hr	Elicitor	Affy ATH1	35650	12773	10080807 27
Harpin Z 1hr	Elicitor	Affy ATH1	32023	11537	10080807 27
Harpin Z 4hr	Elicitor	Affy ATH1	36669	12920	10080807 27
Lipopolysaccharide 1hr	Elicitor	Affy ATH1	33696	11968	10080807 27
Lipopolysaccharide 4hr	Elicitor	Affy	36782	12931	10080807

		ATH1			27
NPP1 (P. parasitica) 1hr	Elicitor	Affy ATH1	31957	11581	10080807 27
NPP1 (P. parasitica) 4hr	Elicitor	Affy ATH1	36873	13037	10080807 27
B. cinerea 18hpi	Fungal pathogen	Affy ATH1	36339	12987	10079674 17
B. cinerea 48hpi	Fungal pathogen	Affy ATH1	34754	12837	10079674 17
E. orontii 12hpi	Fungal pathogen	Affy ATH1	37932	13618	10080314 68
E. orontii 18hpi	Fungal pathogen	Affy ATH1	37675	13561	10080314 68
E. orontii 24hpi	Fungal pathogen	Affy ATH1	37569	13453	10080314 68
E. orontii 2dpi	Fungal pathogen	Affy ATH1	38068	13619	10080314 68
E. orontii 3dpi	Fungal pathogen	Affy ATH1	37847	13556	10080314 68
E. orontii 4dpi	Fungal pathogen	Affy ATH1	38987	14166	10080314 68
E. orontii 5dpi	Fungal pathogen	Affy ATH1	38416	13913	10080314 68
E. orontii 6hpi	Fungal pathogen	Affy ATH1	37968	13798	10080314 68
F. virguliforme	Fungal pathogen	AFGC	19022	8635	10058235 83
P. infestans 12hpi	Fungal pathogen	Affy ATH1	36068	12621	10079660 21
P. infestans 16hpi	Fungal pathogen	AFGC	18701	6742	10058235 34
P. infestans 24hpi	Fungal pathogen	Affy ATH1	35961	12635	10079660 21
P. infestans 6hpi	Fungal pathogen	Affy ATH1	31794	11747	10079660 21
Powdery mildew	Fungal pathogen	AFGC	10159	6995	10058235 49
MYB46 knockout mutant	Mutant	null	37667	28192	null
ABA 1hr (10µM)	Plant hormone	Affy ATH1	25587	13427	10079647 50
ABA 24hr (3µM)	Plant hormone	Affy ATH1	26141	13413	10079673 94
ABA 24hr (30µM)	Plant hormone	Affy ATH1	26647	13665	10079673 94
ABA 30min (10µM)	Plant hormone	Affy ATH1	25726	13432	10079647 50
ABA 3hr (10µM)	Plant hormone	Affy ATH1	24884	13289	10079647 50
Brassinolide 1hr (1µM)	Plant hormone	Affy ATH1	25336	13350	10079660 53
Brassinolide 30min (1µM)	Plant hormone	Affy ATH1	24757	12949	10079660 53
Brassinolide 3hr (1µM)	Plant hormone	Affy ATH1	24854	13102	10079660 53
Brassinolide 3hr (10nm)	Plant hormone	Affy ATH1	28100	14601	10079994 38

GA3 1hr	Plant hormone	Affy ATH1	26370	13844	1007966175
GA3 30min	Plant hormone	Affy ATH1	26183	13632	1007966175
GA3 3hr	Plant hormone	Affy ATH1	25575	13751	1007966175
IAA 1hr	Plant hormone	Affy ATH1	25391	13398	1007965859
IAA 30min	Plant hormone	Affy ATH1	26568	13858	1007965859
IAA 3hr	Plant hormone	Affy ATH1	24999	13405	1007965859
Zeatin 1hr (1µM)	Plant hormone	Affy ATH1	26013	13655	1007966040
Zeatin 30min (1µM)	Plant hormone	Affy ATH1	26428	13771	1007966040
Zeatin 3hr (1µM)	Plant hormone	Affy ATH1	25262	13583	1007966040
Zeatin 3hr (20µM)	Plant hormone	Affy ATH1	41038	14243	1008031453
Cis-jasmone	Signal molecule	AFGC	19978	8051	1005823574
Ethylene 24hr	Signal molecule	Carnegie	9211	1425	1005823581
Ethylene 2hr	Signal molecule	Carnegie	7707	1490	1005823581
Hydrogen peroxide	Signal molecule	AFGC	19097	6902	1005823545
Methyl-jasmonate	Signal molecule	AFGC	20463	8185	1005823574
Methyl-jasmonate 1hr	Signal molecule	Affy ATH1	25371	13337	1007965964
Methyl-jasmonate 30min	Signal molecule	Affy ATH1	25609	13357	1007965964
Methyl-jasmonate 3hr	Signal molecule	Affy ATH1	24213	12795	1007965964
SA analog BTH	Signal molecule	AFGC	9958	6857	1005823548
Salicylic acid	Signal molecule	Affy ATH1	26296	13983	1008080827
TMV infected leaves 3dpi	Viral pathogen	AFGC	17975	6451	1005823504
TMV infected leaves 4dpi	Viral pathogen	AFGC	28467	7860	1005823602
TMV systemic leaves 14dpi	Viral pathogen	AFGC	74561	9321	1005823505
TMV systemic leaves 14dpi	Viral pathogen	AFGC	74561	9321	1005823602

7.2 Normalization values used in the *in silico* expression analysis

Table 7.2: Normalization values used in the *in silico* expression analysis

Stress	Average Mean
ABA 1hr (10µM)	1.029928636
ABA 24hr (3µM)	1.124471728
ABA 24hr (30µM)	1.02221172
ABA 30min (10µM)	1.075517867
ABA 3hr (10µM)	0.949936226
Al-oversupplied roots	0.998972037
B. cinerea 18hpi	0.989673671
B. cinerea 48hpi	0.944121195
Brassinolide 1hr (1µM)	1.169241109
Brassinolide 30min (1µM)	1.15534738
Brassinolide 3hr (1µM)	0.995309406
Brassinolide 3hr (10nm)	0.999626749
Chitin 10min	1.006752947
Chitin 1hr	1.003897586
Chitin 24hr	0.93126143
Chitin 30min	1.004005827
Chitin 3hr	1.007304963
Chitin 6hr	1.006381338
Chitooctaose	0.998669027
Cis-jasmone	1.006915571
Cold-stressed roots 0.5hr	1.019806963
Cold-stressed roots 12hr	0.946667899
Cold-stressed roots 1hr	1.067822182
Cold-stressed roots 24hr	0.944969256
Cold-stressed roots 3hr	0.961168544
Cold-stressed roots 6hr	1.011113918
Cold-stressed shoots 0.5hr	1.044150122
Cold-stressed shoots 12hr	0.960081539
Cold-stressed shoots 1hr	1.011422336
Cold-stressed shoots 24hr	0.882721539
Cold-stressed shoots 3hr	0.958680332
Cold-stressed shoots 6hr	0.901947021
Drought-stressed roots 0.25hr	0.933859721
Drought-stressed roots 0.5hr	0.957730983
Drought-stressed roots 12hr	0.967972972
Drought-stressed roots 1hr	0.992522846
Drought-stressed roots 24hr	0.952608644
Drought-stressed roots 3hr	0.980098624
Drought-stressed roots 6hr	0.941189697
Drought-stressed shoots 0.25hr	0.902370675
Drought-stressed shoots 0.5hr	1.006043582

Drought-stressed shoots 12hr	0.977077313
Drought-stressed shoots 1hr	1.008714621
Drought-stressed shoots 24hr	1.058497668
Drought-stressed shoots 3hr	0.948096728
Drought-stressed shoots 6hr	0.908067936
E. orontii 12hpi	0.995222887
E. orontii 18hpi	0.909994465
E. orontii 24hpi	0.973019338
E. orontii 2dpi	0.945669013
E. orontii 3dpi	1.05942434
E. orontii 4dpi	0.962828577
E. orontii 5dpi	1.047683505
E. orontii 6hpi	0.945623664
EF-Tu 30min	1.002450809
EF-Tu 60min	0.916478509
Ethylene 24hr	1.006442434
Ethylene 2hr	1.00764587
F. virguliforme	0.99738169
Flg22 (P. syringae) 1hr	0.987848941
Flg22 (P. syringae) 4hr	1.076063717
GA3 1hr	1.050952455
GA3 30min	1.018136832
GA3 3hr	1.042496669
Harpin Z 1hr	0.964035864
Harpin Z 4hr	0.990336022
Hydrogen peroxide	0.999532649
IAA 1hr	1.014039255
IAA 30min	1.013317943
IAA 3hr	0.990946379
Inflorescence vs. shoot apex, vegetative	0.984552927
Inflorescence vs. young leaves	1.020606511
Lipopolysaccharide 1hr	1.00683412
Lipopolysaccharide 4hr	1.024930519
Methyl-jasmonate	1.012498299
Methyl-jasmonate 1hr	1.032036824
Methyl-jasmonate 30min	1.081485663
Methyl-jasmonate 3hr	0.931921997
MYB46 knockout mutant	0.999281406
NPP1 (P. parasitica) 1hr	0.995217899
NPP1 (P. parasitica) 4hr	1.00474784
Osmotic-stressed roots 0.5hr	1.141933882
Osmotic-stressed roots 12hr	1.038514718
Osmotic-stressed roots 1hr	1.046510505
Osmotic-stressed roots 24hr	1.050450918
Osmotic-stressed roots 3hr	0.973804249

Osmotic-stressed roots 6hr	0.98254354
Osmotic-stressed shoots 0.5hr	1.059369313
Osmotic-stressed shoots 12hr	0.983319665
Osmotic-stressed shoots 1hr	0.99986847
Osmotic-stressed shoots 24hr	0.9157276
Osmotic-stressed shoots 3hr	0.967146052
Osmotic-stressed shoots 6hr	0.914636359
P. infestans 12hpi	0.950186492
P. infestans 16hpi	1.002260107
P. infestans 24hpi	0.980694667
P. infestans 6hpi	1.018845857
P. syringae pv. maculicola 16hpi	1.040462528
P. syringae pv. maculicola 24hpi	1.02827705
P. syringae pv. maculicola 48hpi	1.023318987
P. syringae pv. maculicola 4hpi	0.985063433
P. syringae pv. maculicola 8hpi	0.937498493
P. syringae pv. maculicola avrRpt2- 16hpi	1.022855011
P. syringae pv. maculicola avrRpt2- 24hpi	0.918054857
P. syringae pv. maculicola avrRpt2- 48hpi	0.898612997
P. syringae pv. maculicola avrRpt2- 4hpi	0.893115761
P. syringae pv. maculicola avrRpt2- 8hpi	0.982389276
P. syringae pv. phaseolicola 24hpi	0.942079051
P. syringae pv. phaseolicola 2hpi	0.98412671
P. syringae pv. phaseolicola 6hpi	0.95021288
P. syringae pv. tomato 24hpi	0.906152127
P. syringae pv. tomato 2hpi	1.005435952
P. syringae pv. tomato 6hpi	0.996612966
P. syringae pv. tomato avrRpm1 24hpi	0.917836317
P. syringae pv. tomato avrRpm1 2hpi	0.989965283
P. syringae pv. tomato avrRpm1 6hpi	0.973177888
P. syringae pv. tomato hrcC- 24hpi	0.931691462
P. syringae pv. tomato hrcC- 2hpi	0.981146467
P. syringae pv. tomato hrcC- 6hpi	0.960011057
Pb-oversupplied (25ppm) leaves	1.157926421
Pb-oversupplied (25ppm) roots	1.0676348
Pb-oversupplied (50ppm) leaves	1.082582148
Pb-oversupplied (50ppm) roots	1.194321711
Powdery mildew	0.998633514
SA analog BTH	1.001522737
Salicylic acid	0.917559219
Salt-stressed roots 0.5hr	1.032310988
Salt-stressed roots 12hr	1.015427485
Salt-stressed roots 1hr	0.996621012
Salt-stressed roots 24hr	1.034556571
Salt-stressed roots 3hr	0.93523461

Salt-stressed roots 6hr	1.007103768
Salt-stressed shoots 0.5hr	1.003576722
Salt-stressed shoots 12hr	1.031519662
Salt-stressed shoots 1hr	1.029614473
Salt-stressed shoots 24hr	0.973780106
Salt-stressed shoots 3hr	1.034040323
Salt-stressed shoots 6hr	1.017289978
TMV infected leaves 3dpi	1.001661145
TMV infected leaves 4dpi	1.001876639
TMV systemic leaves 14dpi	0.994777634
X. campestris	0.99950798
Zeatin 1hr (1µM)	1.037167121
Zeatin 30min (1µM)	1.020463265
Zeatin 3hr (1µM)	0.993029511
Zeatin 3hr (20µM)	1.008343724
Zn-deficient roots	1.040019229
Zn-deficient shoots	0.990849814
Zn-oversupplied roots 2hr	0.970471906
Zn-oversupplied roots 8hr	1.066044167
Zn-oversupplied shoots 8hr	1.219493619
Zn-resupplied roots 2hr vs. deficient Zn	0.95514562
Zn-resupplied roots 2hr vs. suffficient Zn	0.993369542
Zn-resupplied shoots 8hr vs. deficient Zn	1.078706238
Zn-resupplied shoots 8hr vs. suffficient Zn	1.068837555

7.3 Similarity trees of MEME predicted motifs

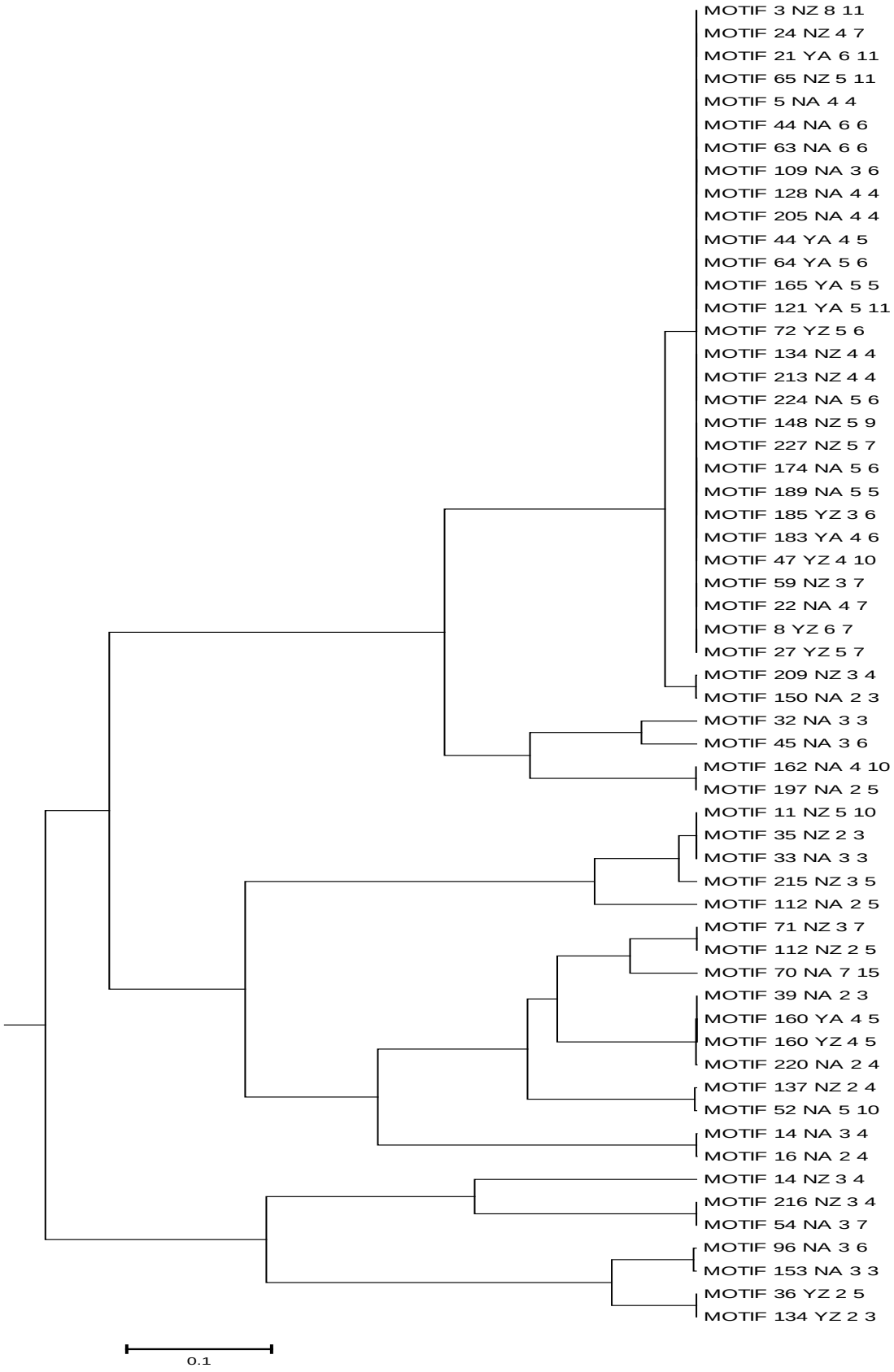


Figure 7.1: Similarity tree of Zn-Deficiency predicted CREs.

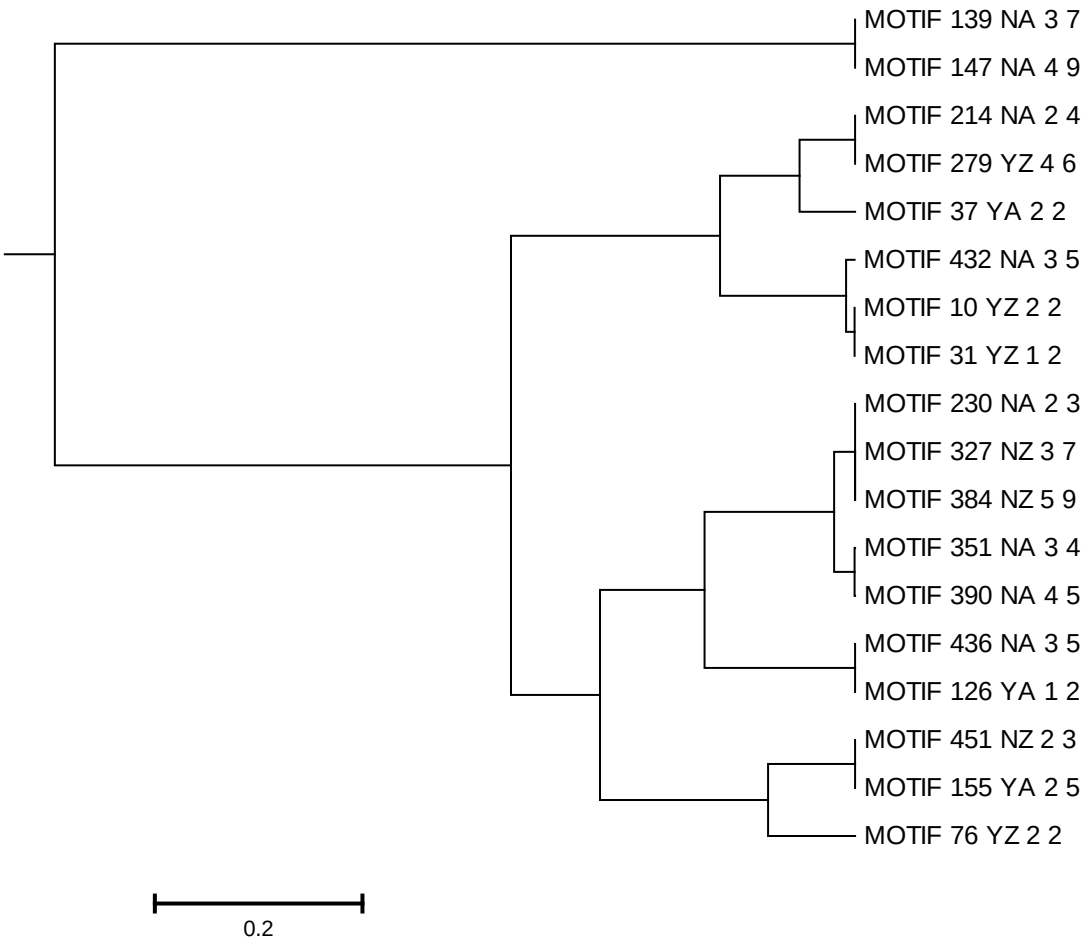


Figure 7.2: Similarity tree of Zn-Oversupply predicted CREs.

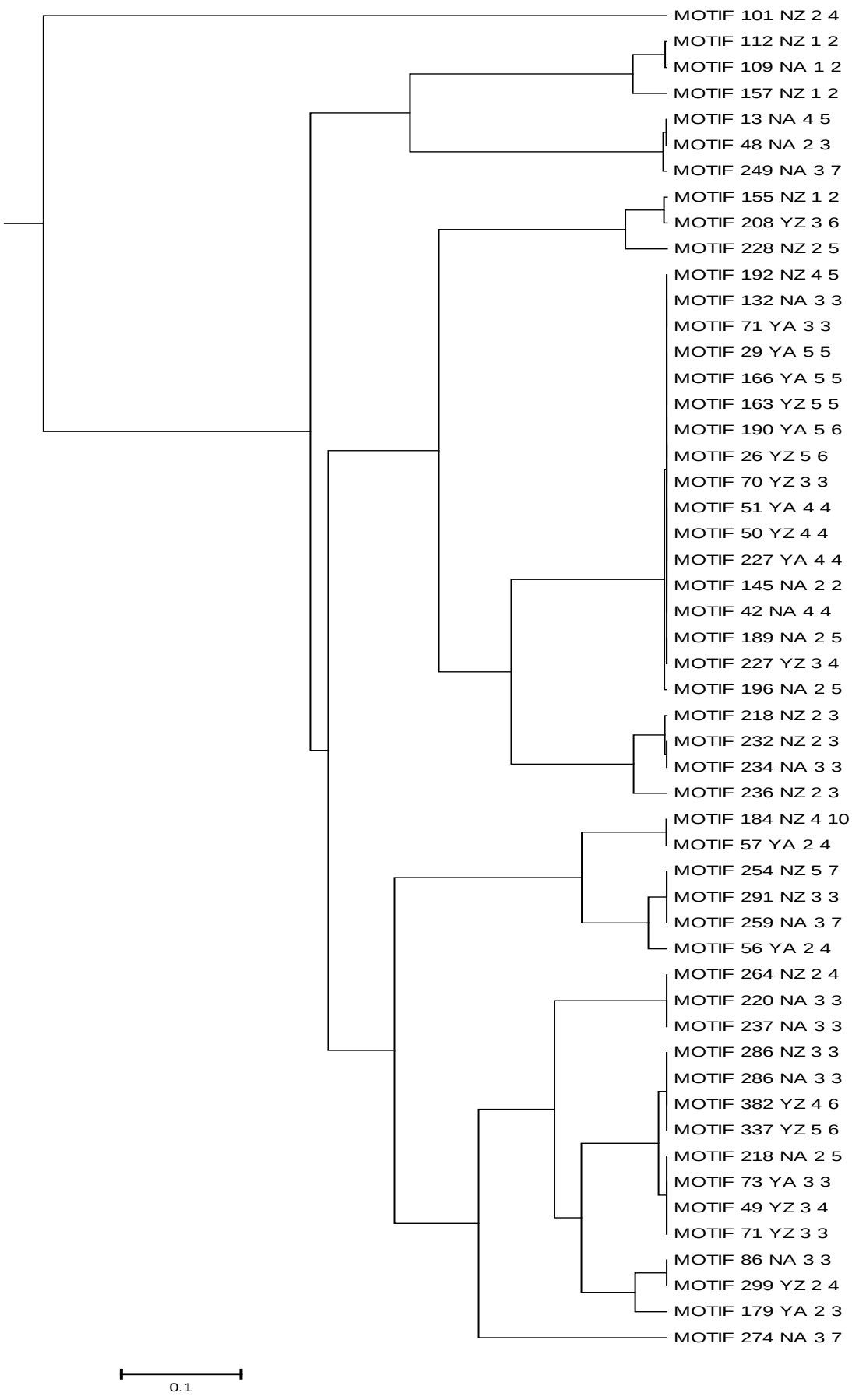


Figure 7.3: Similarity tree of Zn-Deficiency vs resupplied predicted CREs.

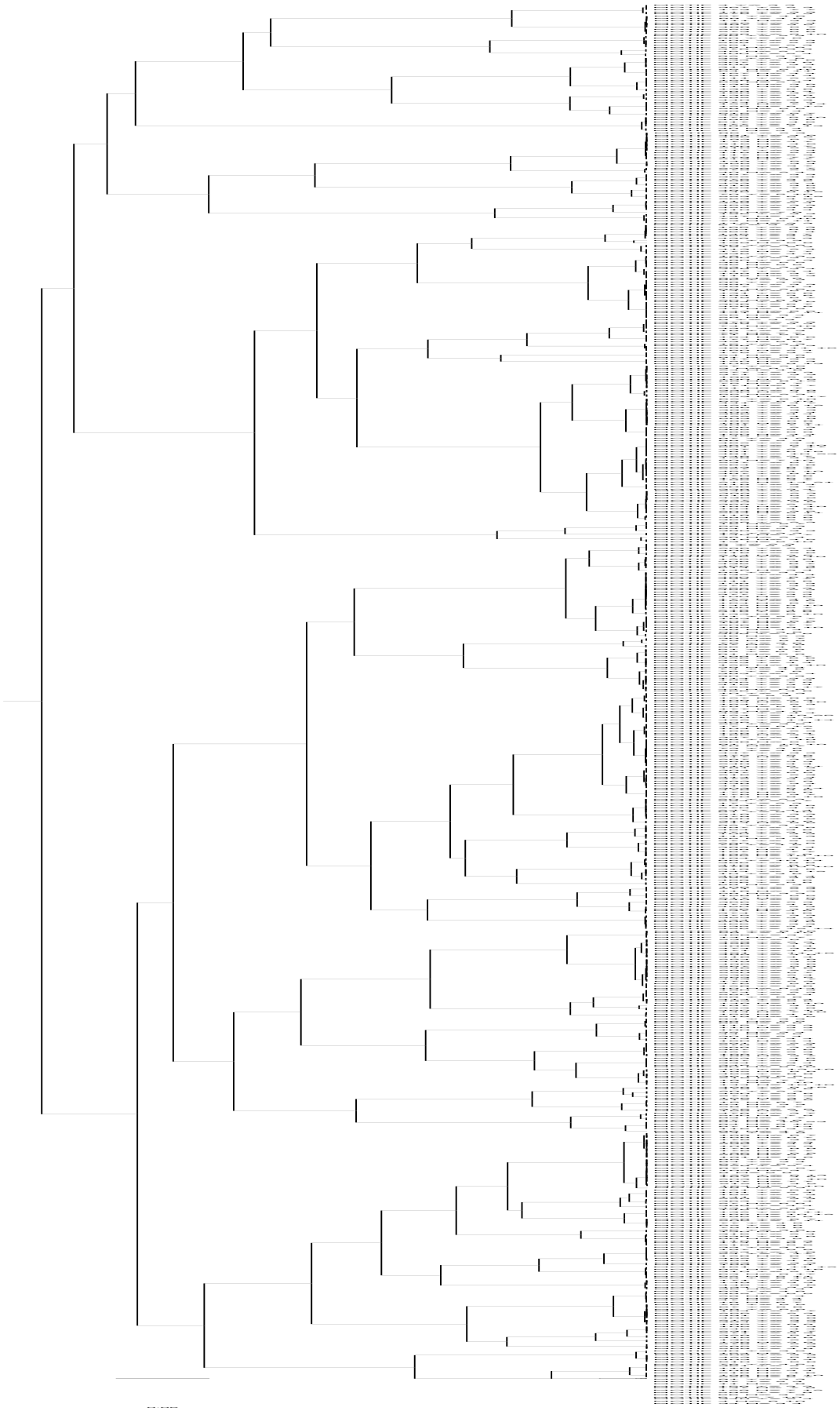


Figure 7.4: Similarity tree of Pb-Oversupply predicted CREs. The high number of elements does not allow element name displaying.

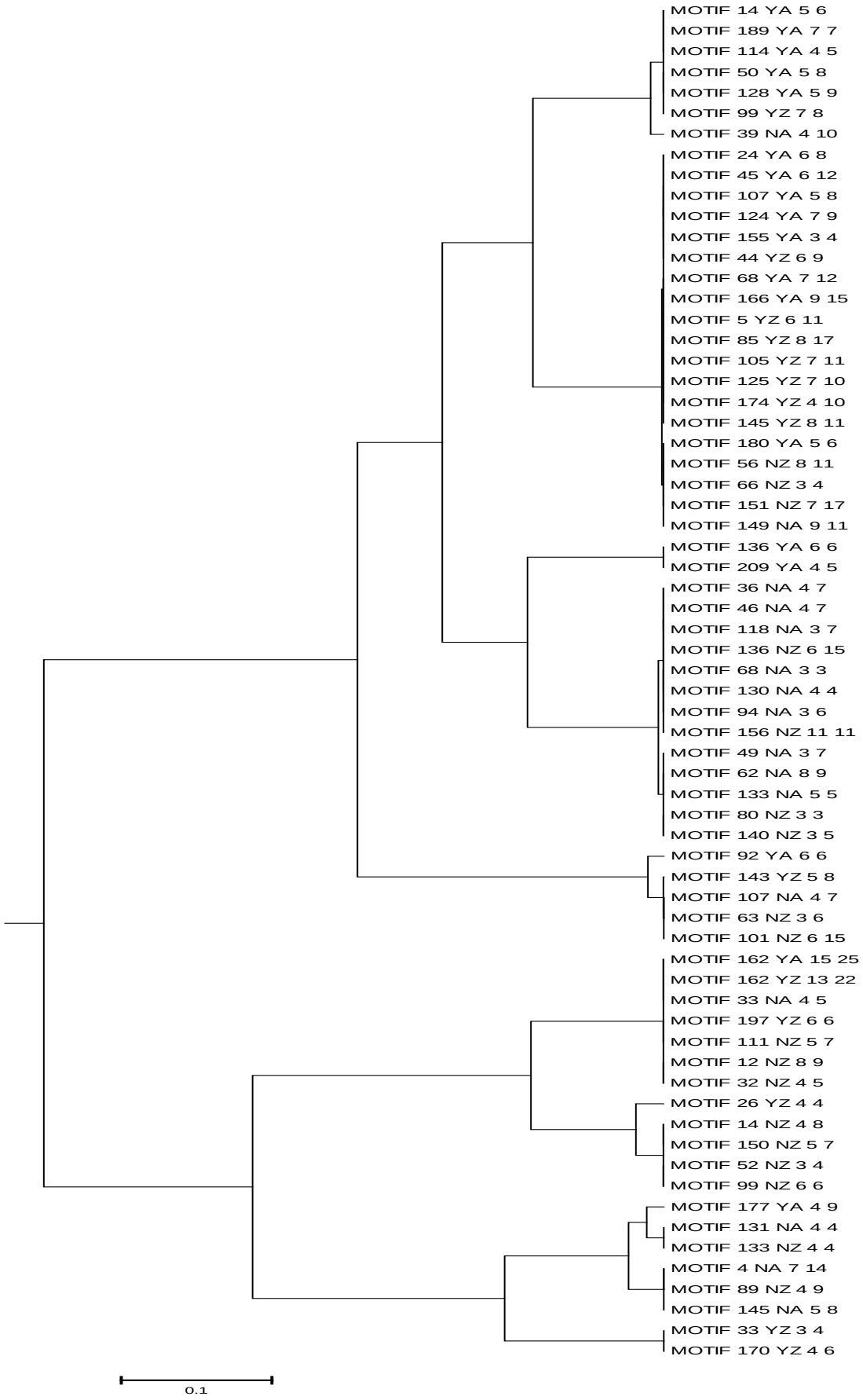


Figure 7.5: Similarity tree of EF-Tu predicted CREs.

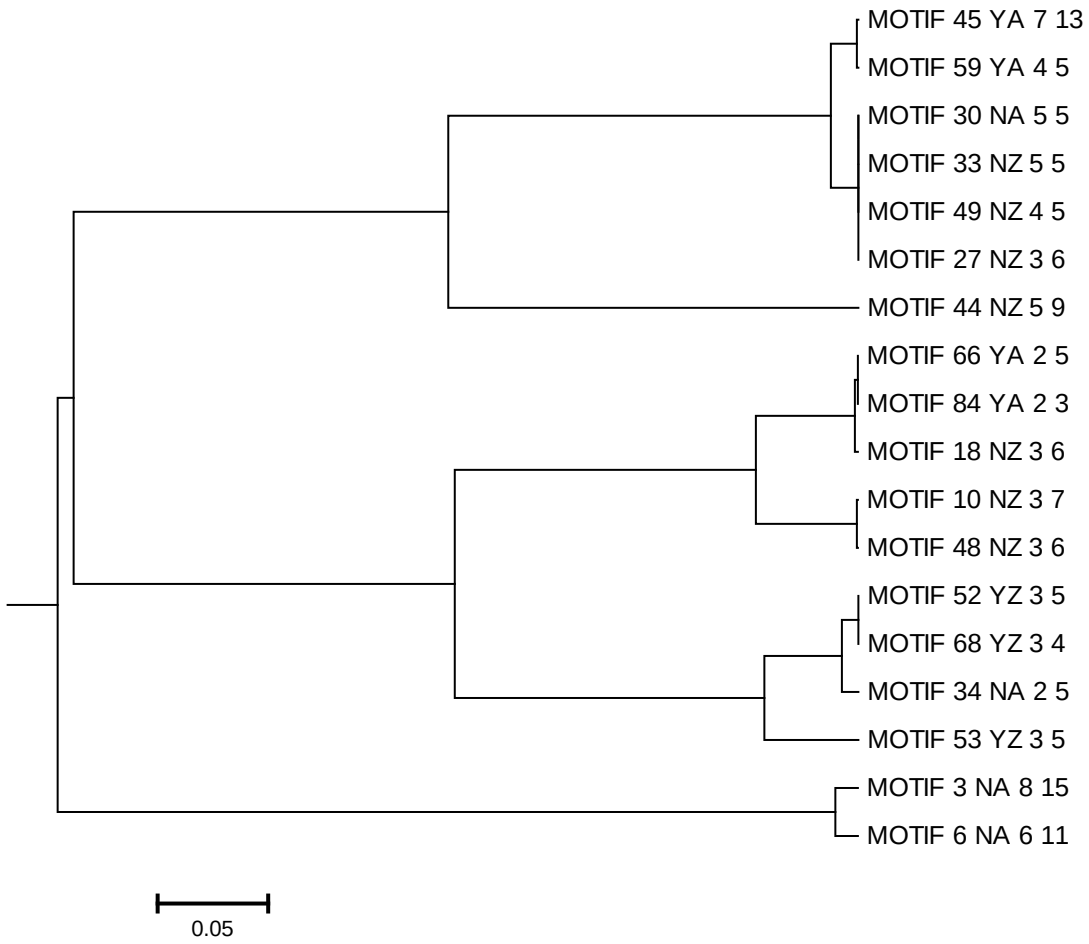


Figure 7.6: Similarity tree of Chitooctaose predicted CREs.

7.4 Combinatorial element spacer lengths

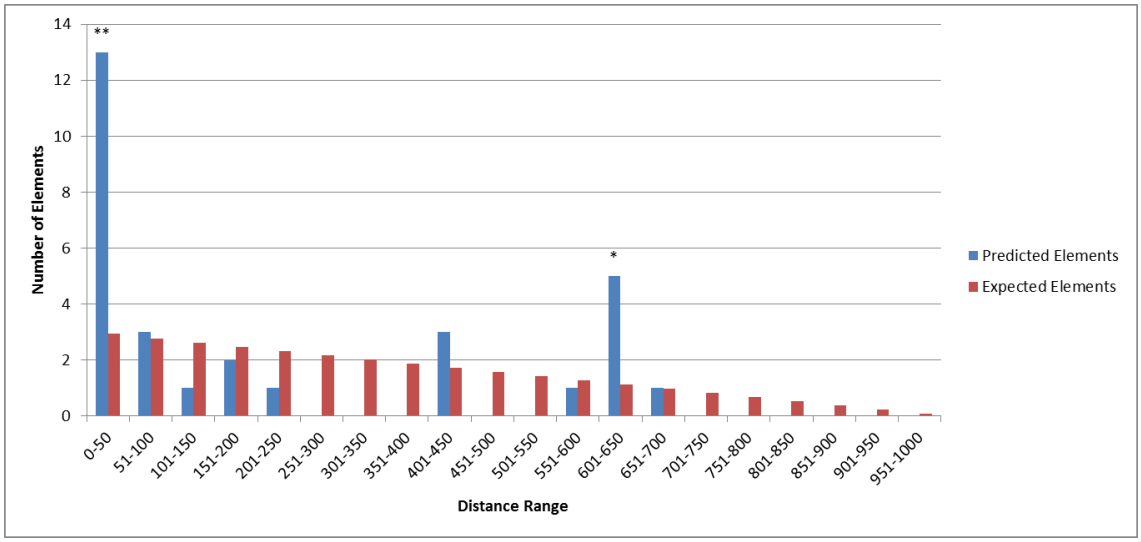


Figure 7.7: Distribution of combinatorial elements putatively responsive to the stress Chitooctaose according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability ≤0.01 and ** a probability ≤0.001 of observing exactly the number of predicted elements given the expected ones.

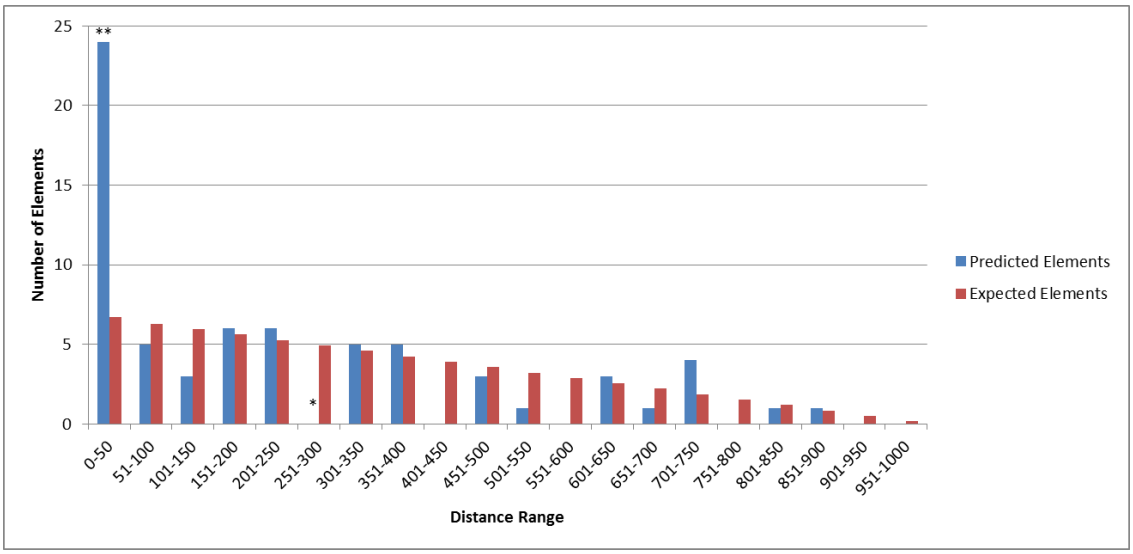


Figure 7.8: Distribution of combinatorial elements putatively responsive to the stress EF-Tu 30min according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability ≤0.01 and ** a probability ≤0.001 of observing exactly the number of predicted elements given the expected ones.

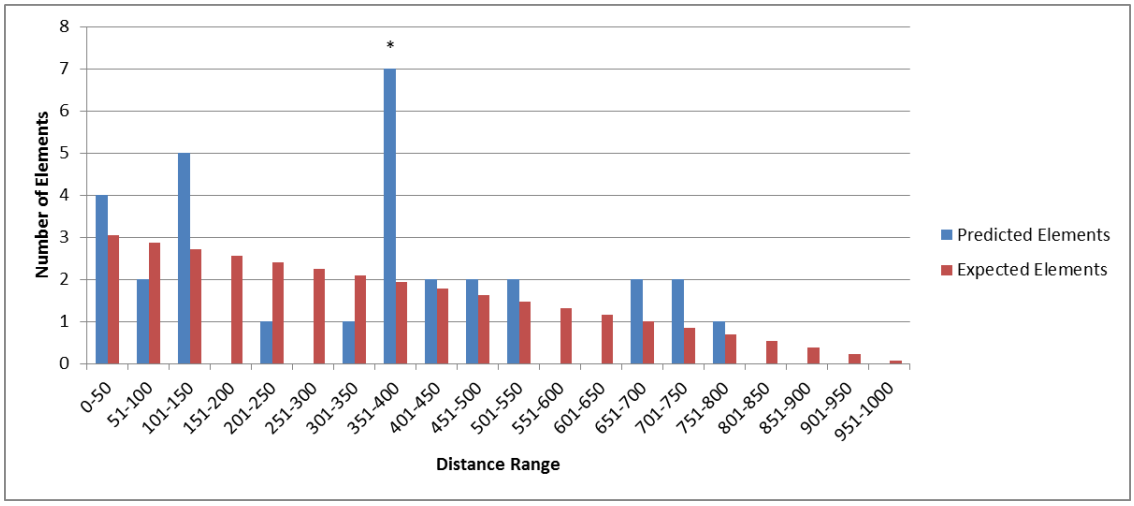


Figure 7.9: Distribution of combinatorial elements putatively responsive to the stress Pb 25ppm roots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

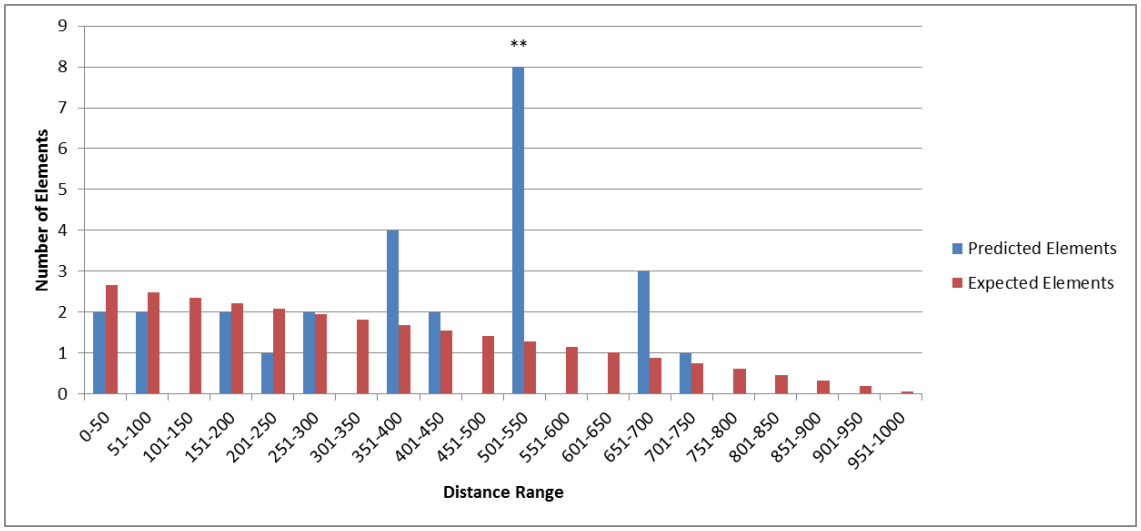


Figure 7.10: Distribution of combinatorial elements putatively responsive to the stress Pb 50ppm leaves according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

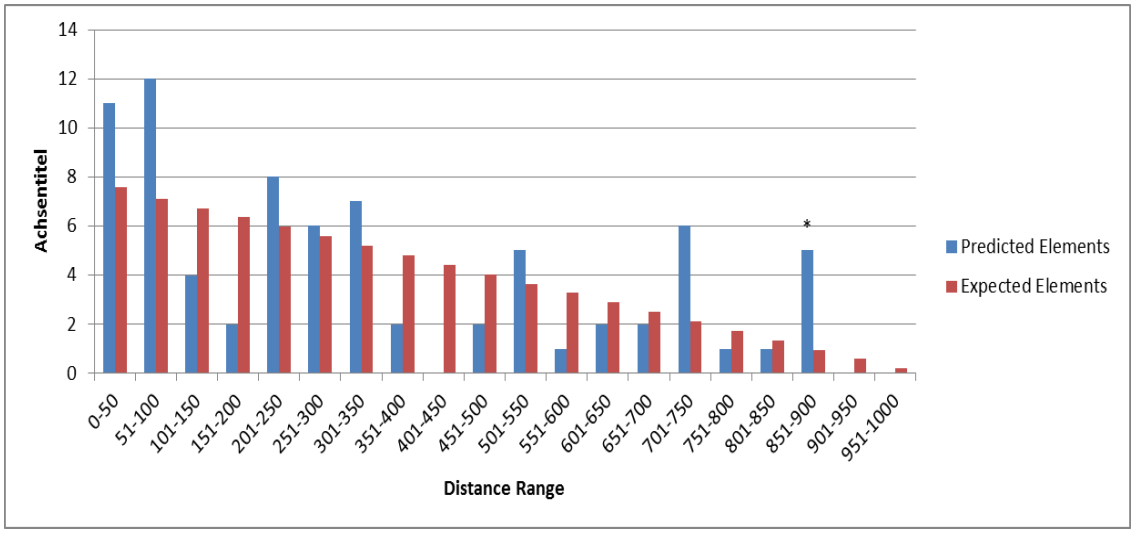


Figure 7.11: Distribution of combinatorial elements putatively responsive to the stress Pb 50ppm roots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

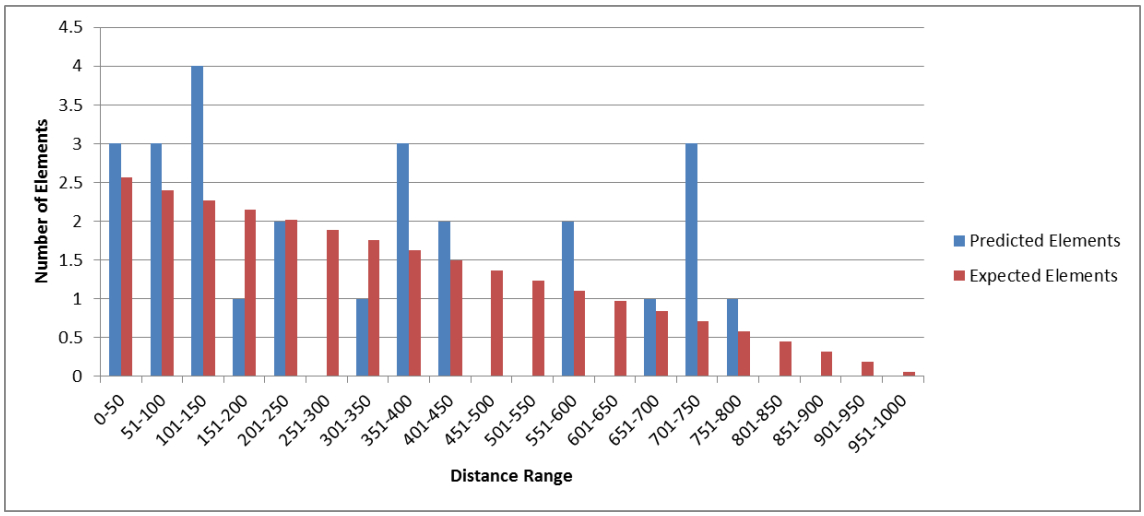


Figure 7.12: Distribution of combinatorial elements putatively responsive to the stress Zn-deficient roots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns.

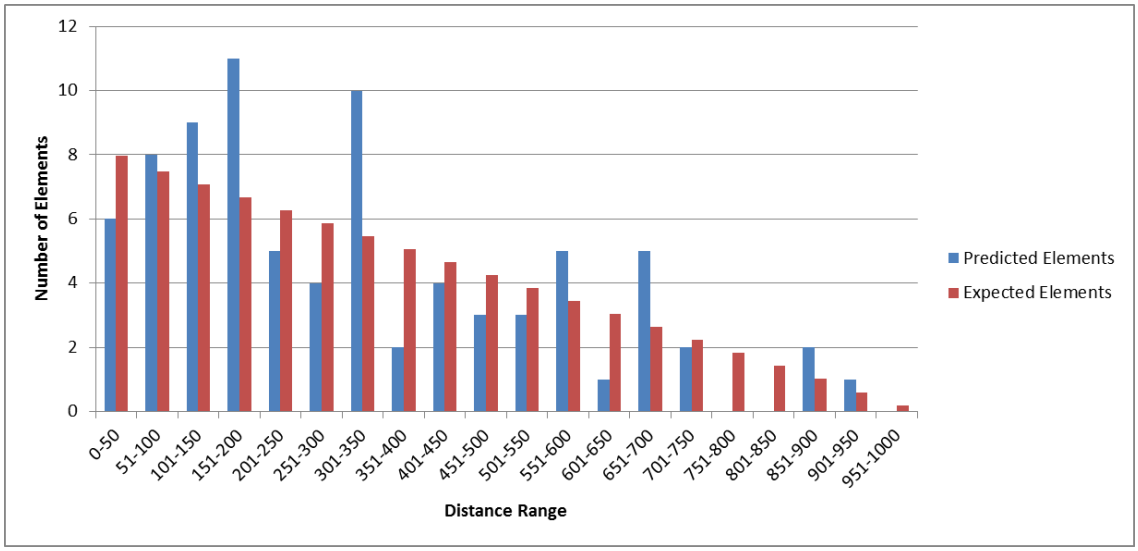


Figure 7.13: Distribution of combinatorial elements putatively responsive to the stress Zn-deficient shoots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns.

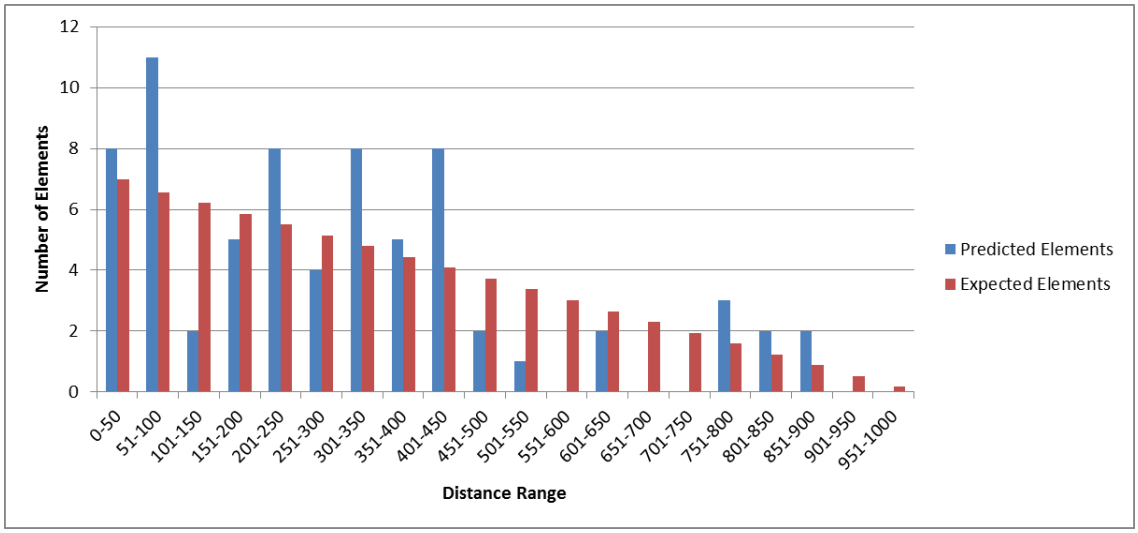


Figure 7.14: Distribution of combinatorial elements putatively responsive to the stress Zn-resupplied 2h roots vs. deficient Zn shoots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns.

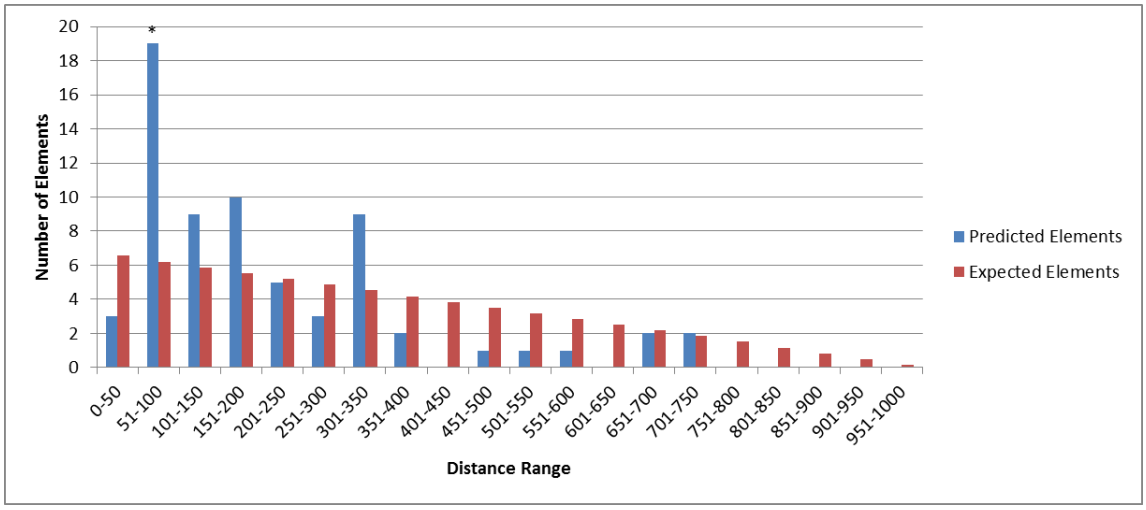


Figure 7.15: Distribution of combinatorial elements putatively responsive to the stress Zn-resupplied 2h roots vs. sufficient Zn shoots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

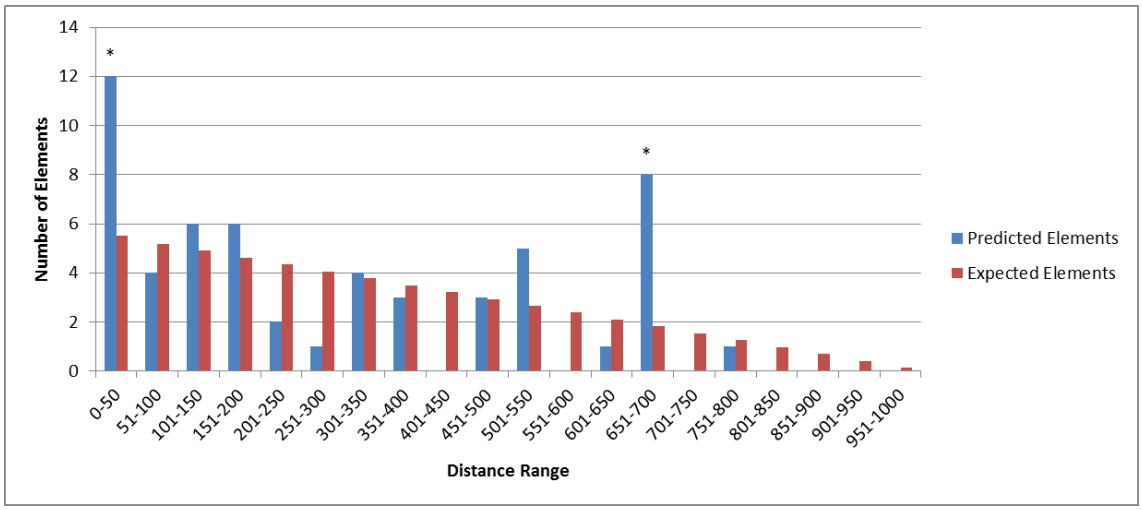


Figure 7.16: Distribution of combinatorial elements putatively responsive to the stress Zn-resupplied 8h shoots vs. sufficient Zn shoots according to their spacer lengths. The number of predicted elements having a spacer length within a given range (x axis) is represented as blue columns. The number of randomly expected elements at a given range is represented by red columns. An * indicates a probability $\leq$0.01 and ** a probability $\leq$0.001 of observing exactly the number of predicted elements given the expected ones.

7.5 Orientation frequencies among predicted combinatorial element sets

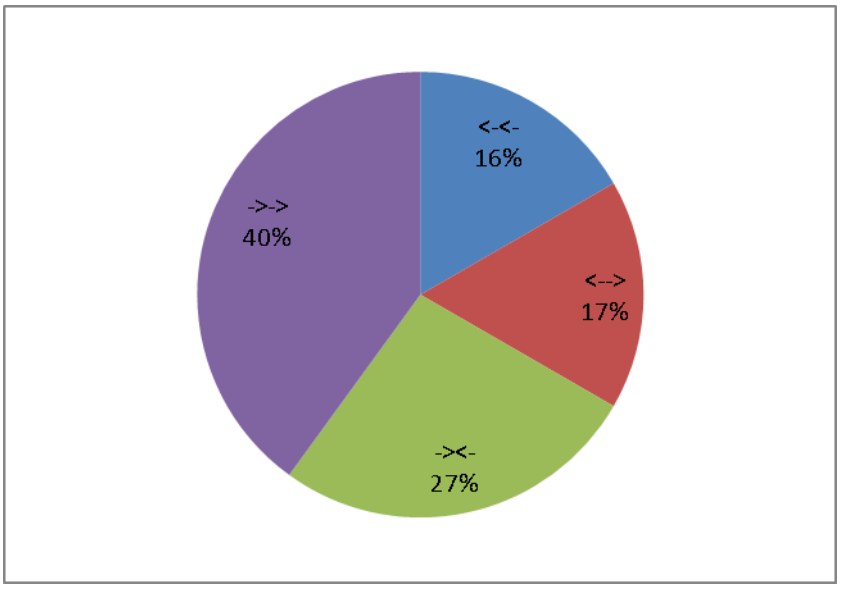


Figure 7.17: Orientations frequency of putatively Chitooctaose responsive combinatorial elements.

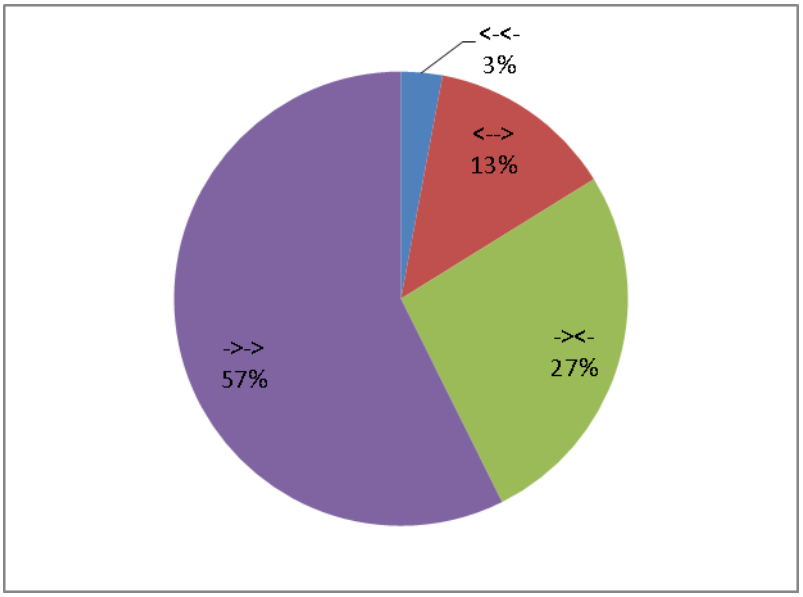


Figure 7.18: Orientations frequency of putatively EFTu30min responsive combinatorial elements.

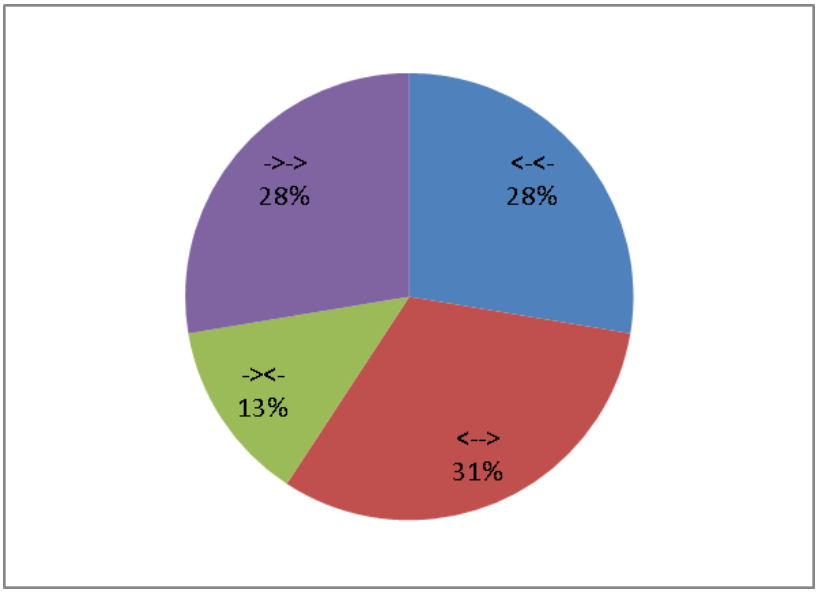


Figure 7.19: Orientations frequency of putatively EFTu60min responsive combinatorial elements.

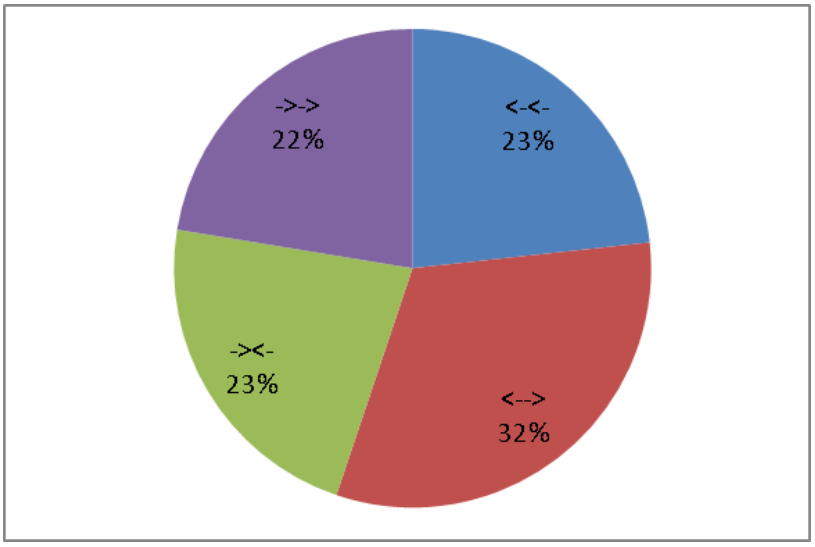


Figure 7.20: Orientations frequency of putatively Flg22 4h responsive combinatorial elements.

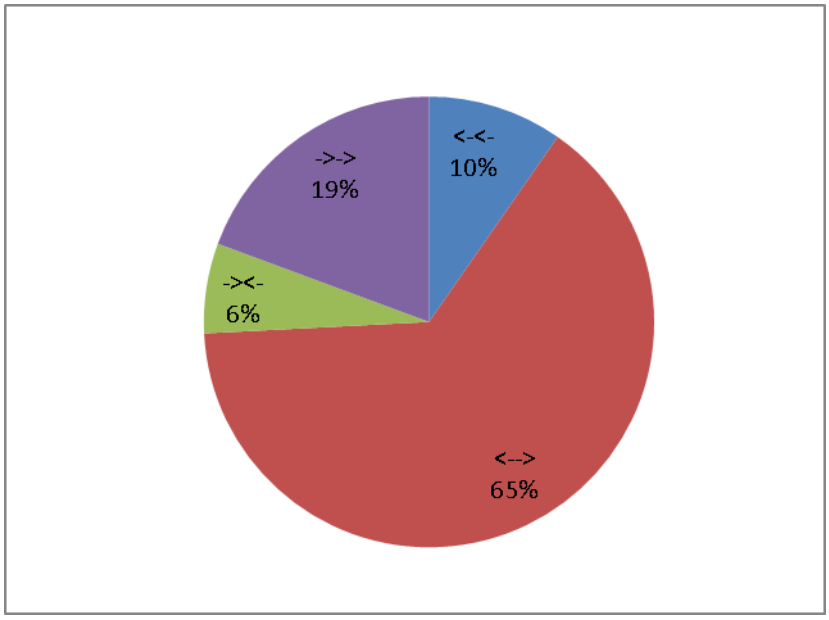


Figure 7.21: Orientations frequency of putatively Pb25ppm roots responsive combinatorial elements.

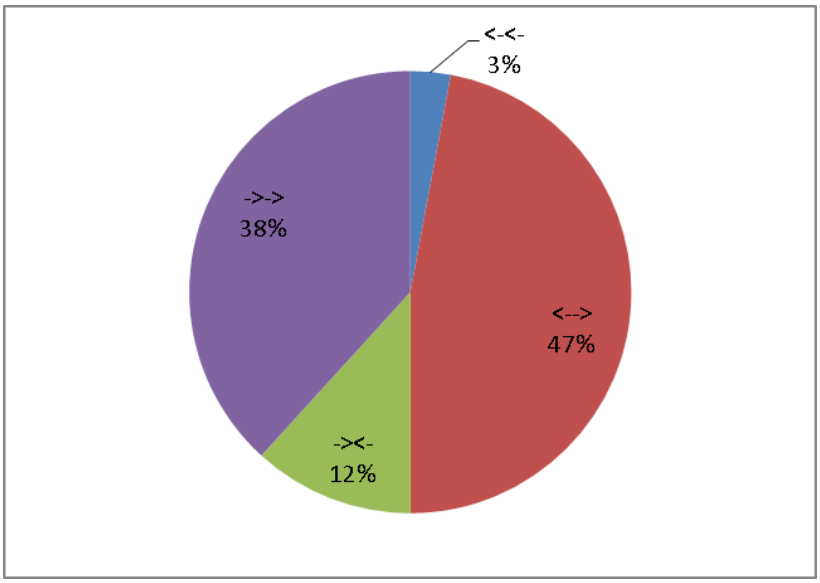


Figure 7.22: Orientations frequency of putatively Pb50ppm leaves responsive combinatorial elements.

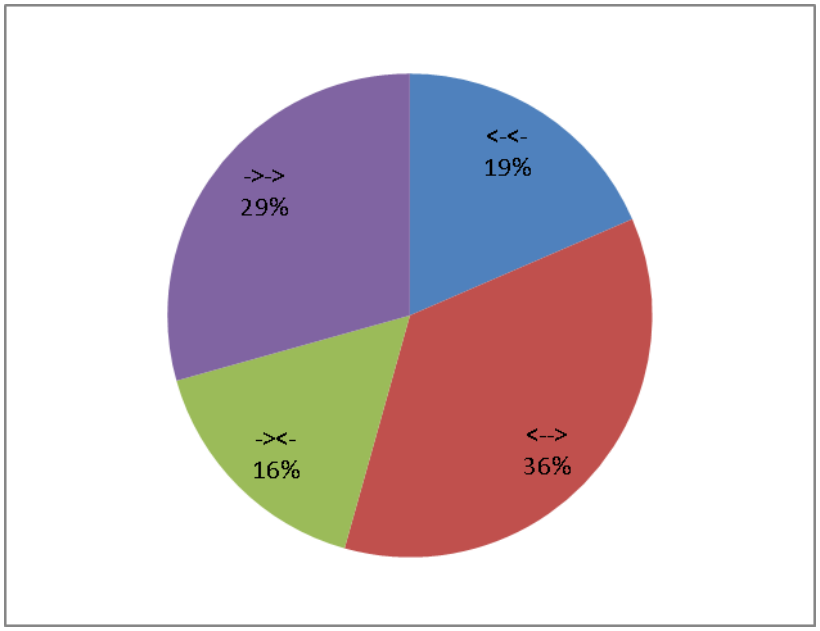


Figure 7.23: Orientations frequency of putatively Pb50ppm roots responsive combinatorial elements.

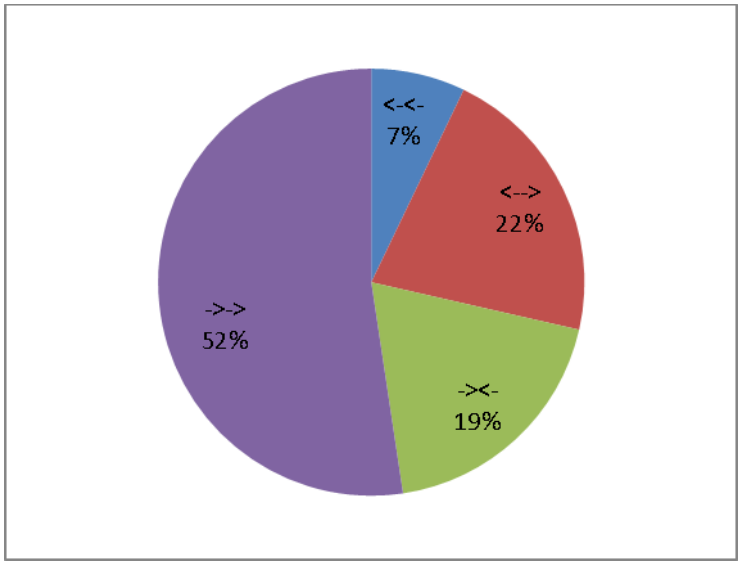


Figure 7.24: Orientations frequency of putatively Zn-deficiency roots responsive combinatorial elements.

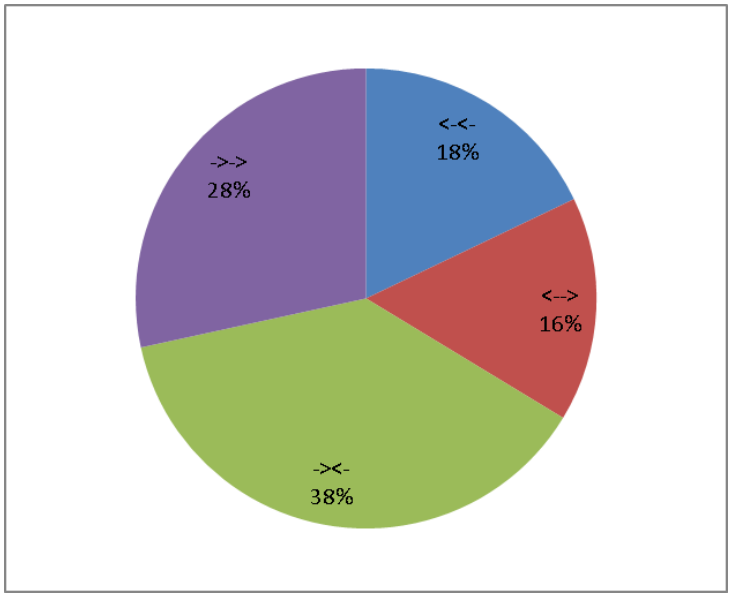


Figure 7.25: Orientations frequency of putatively Zn-deficiency shoots responsive combinatorial elements.

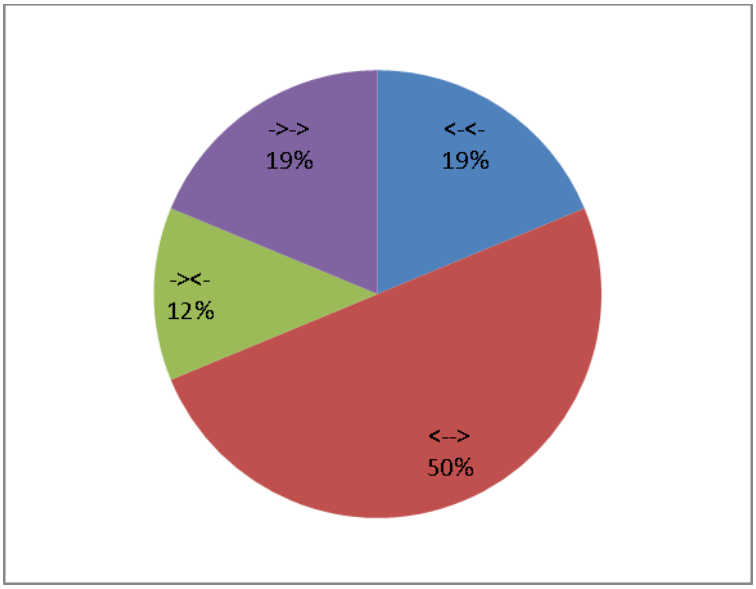


Figure 7.26: Orientations frequency of putatively Zn-oversupply 2h roots responsive combinatorial elements.

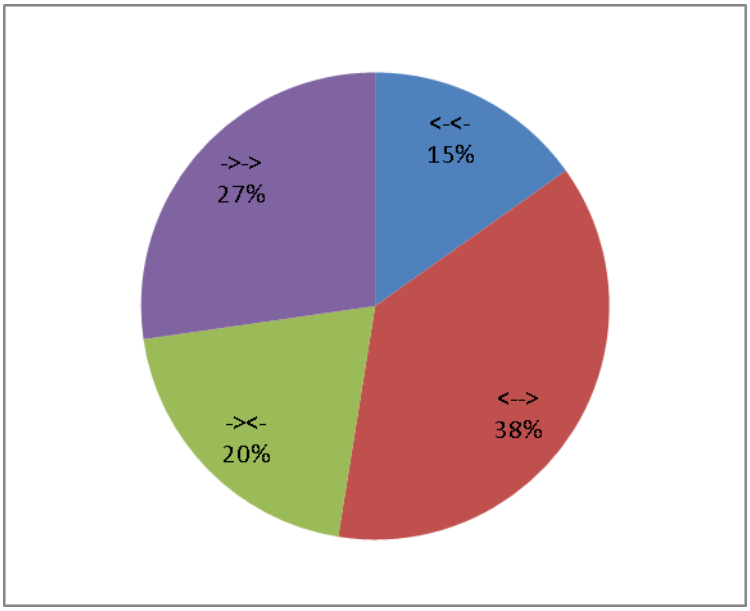


Figure 7.27: Orientations frequency of putatively Zn-resupplied roots 2h vs. deficiency responsive combinatorial elements.

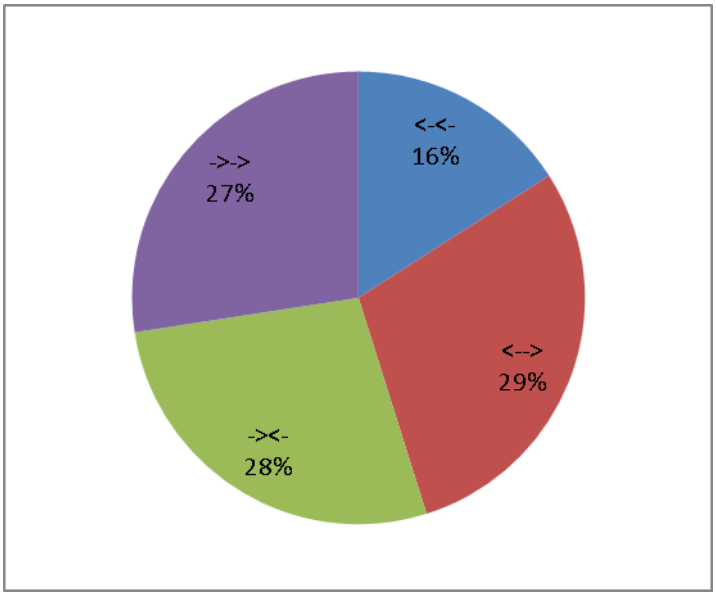


Figure 7.28: Orientations frequency of putatively Zn-resupplied roots 2h vs. sufficient Zn responsive combinatorial elements.

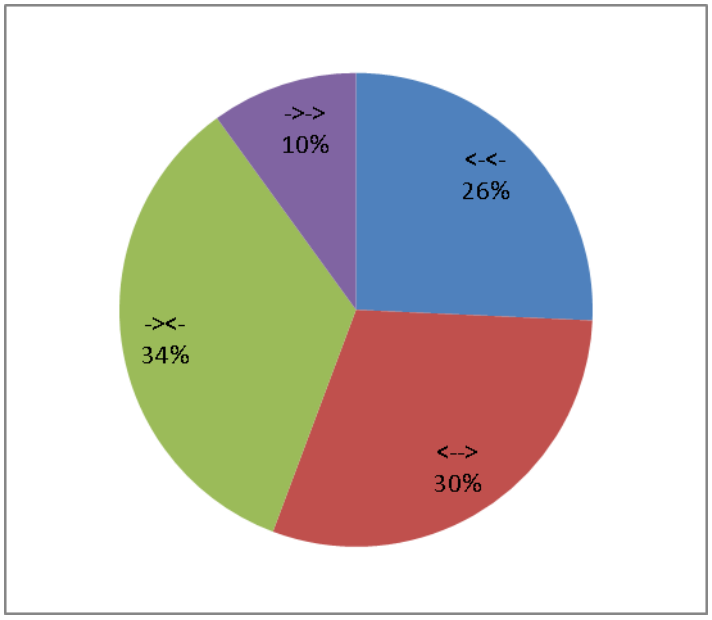


Figure 7.29: Orientations frequency of putatively Zn-resupplied shoots 8h vs. sufficient Zn responsive combinatorial elements.

7.6 Combinatorial element distances to the TSS

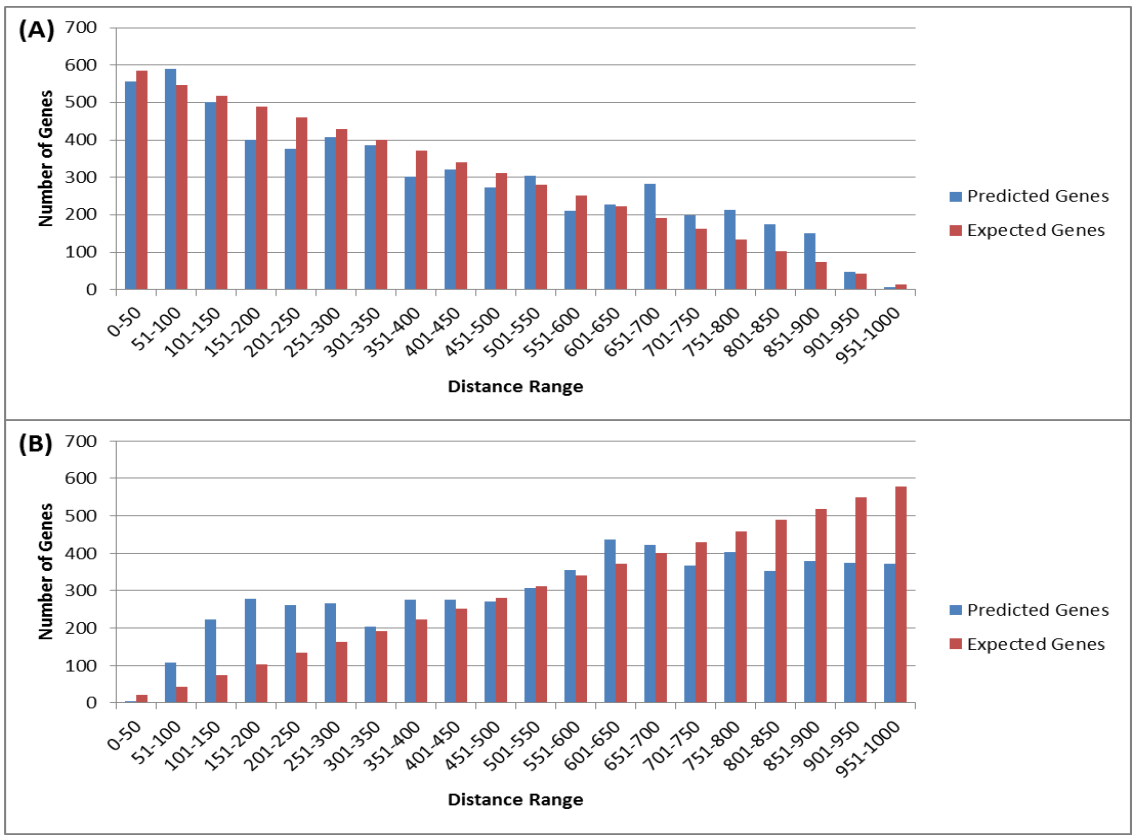


Figure 7.30: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Flg22 4h.

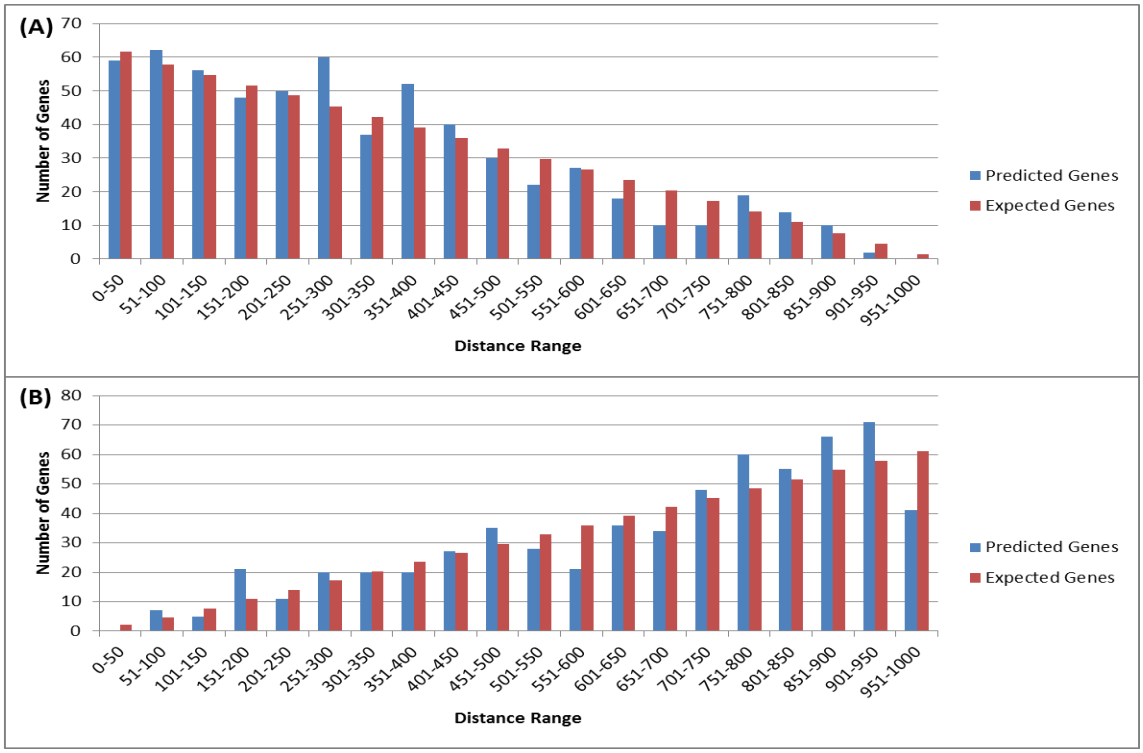


Figure 7.31: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Pb 25ppm roots.

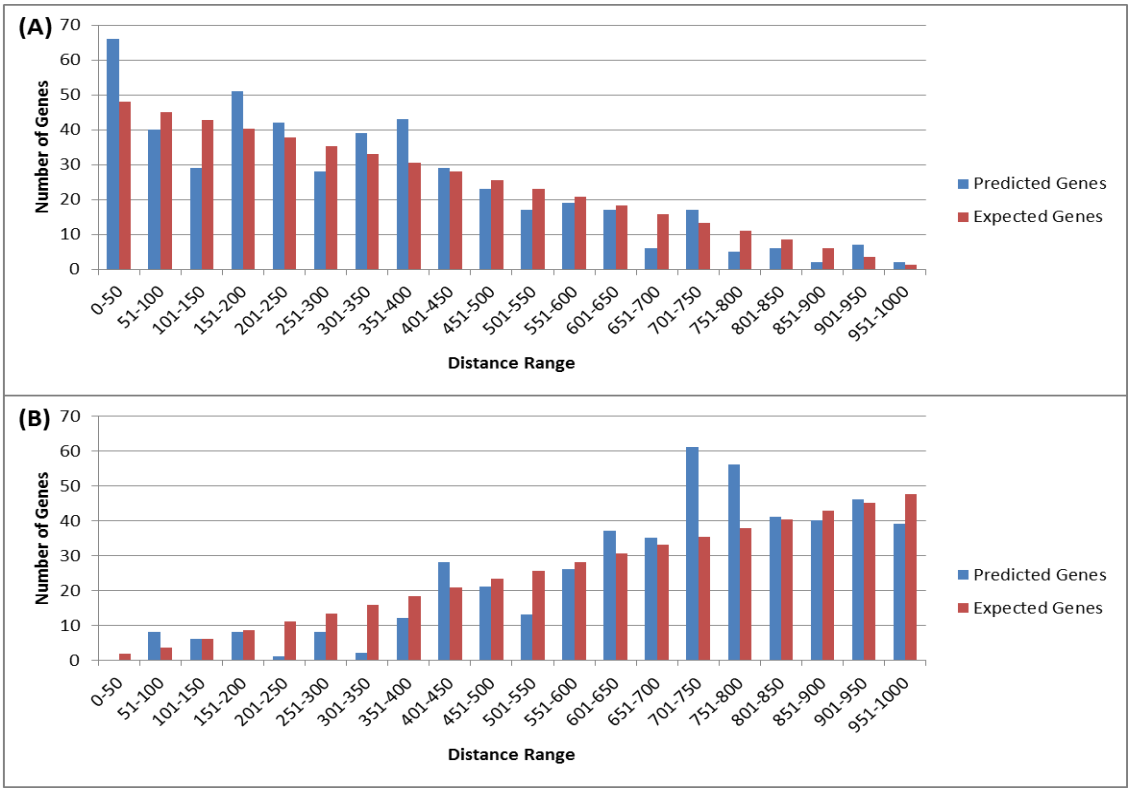


Figure 7.32: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Pb 50ppm leaves.

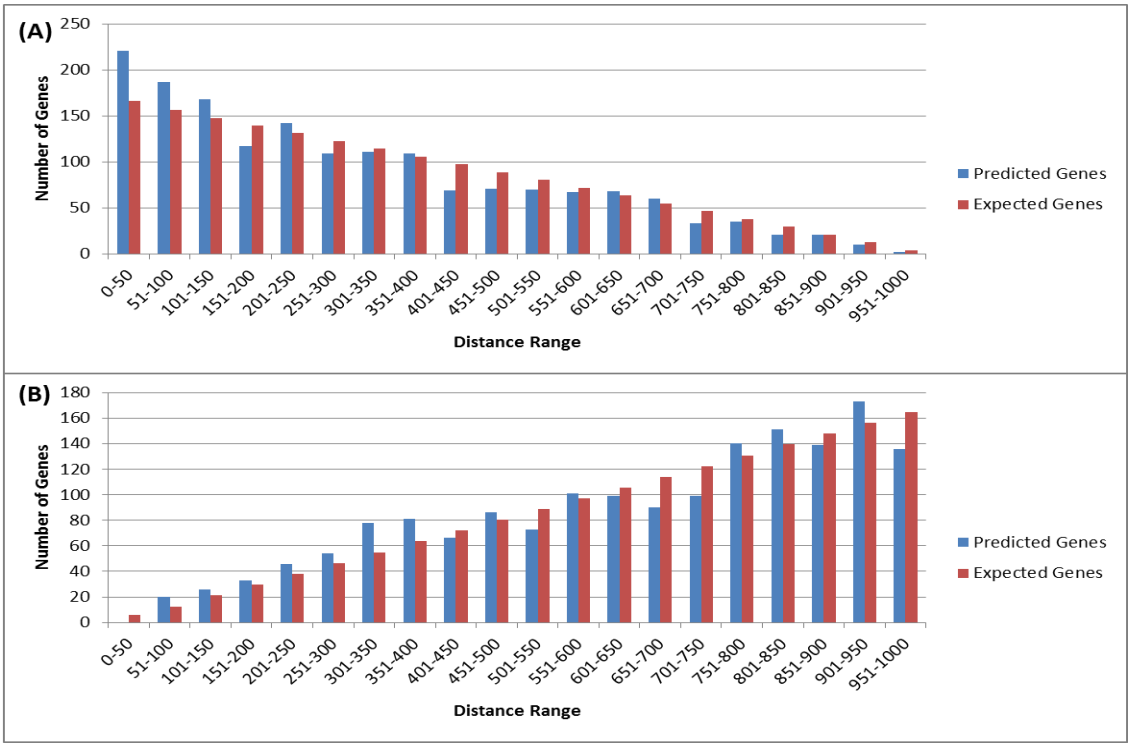


Figure 7.33: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Pb 50ppm roots.

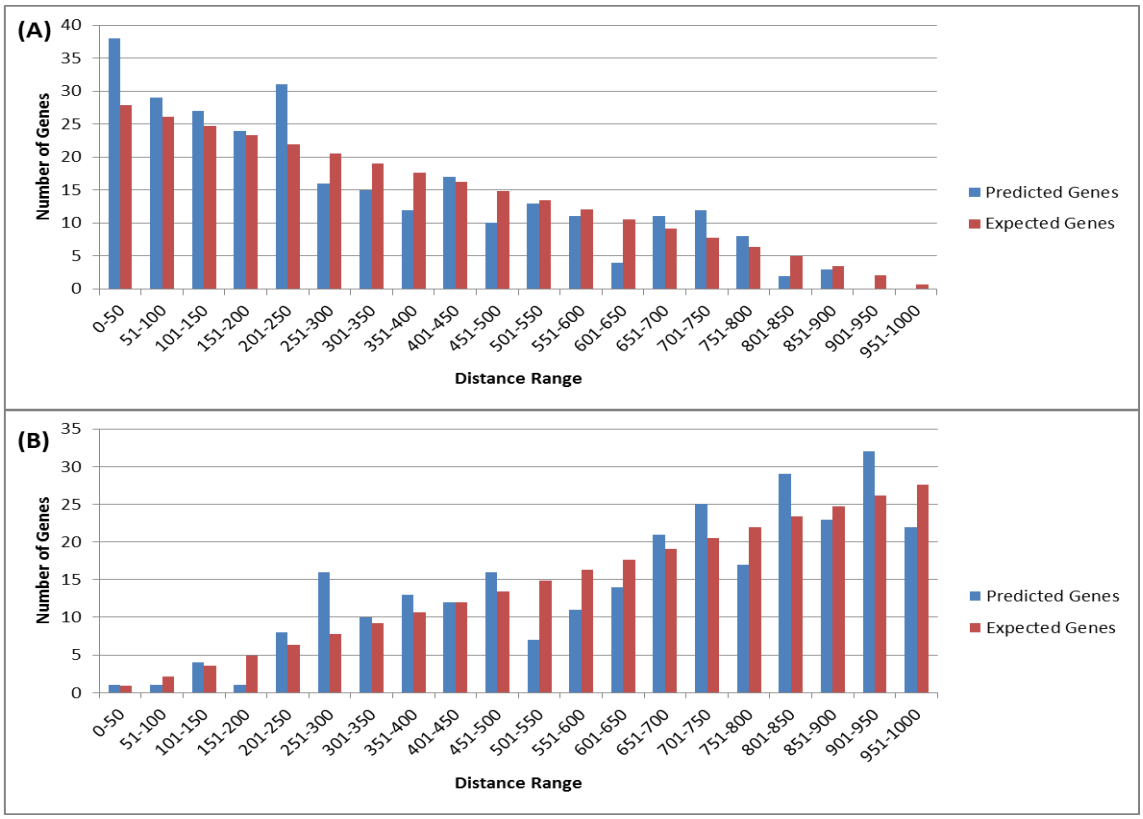


Figure 7.34: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Zn-deficiency roots.

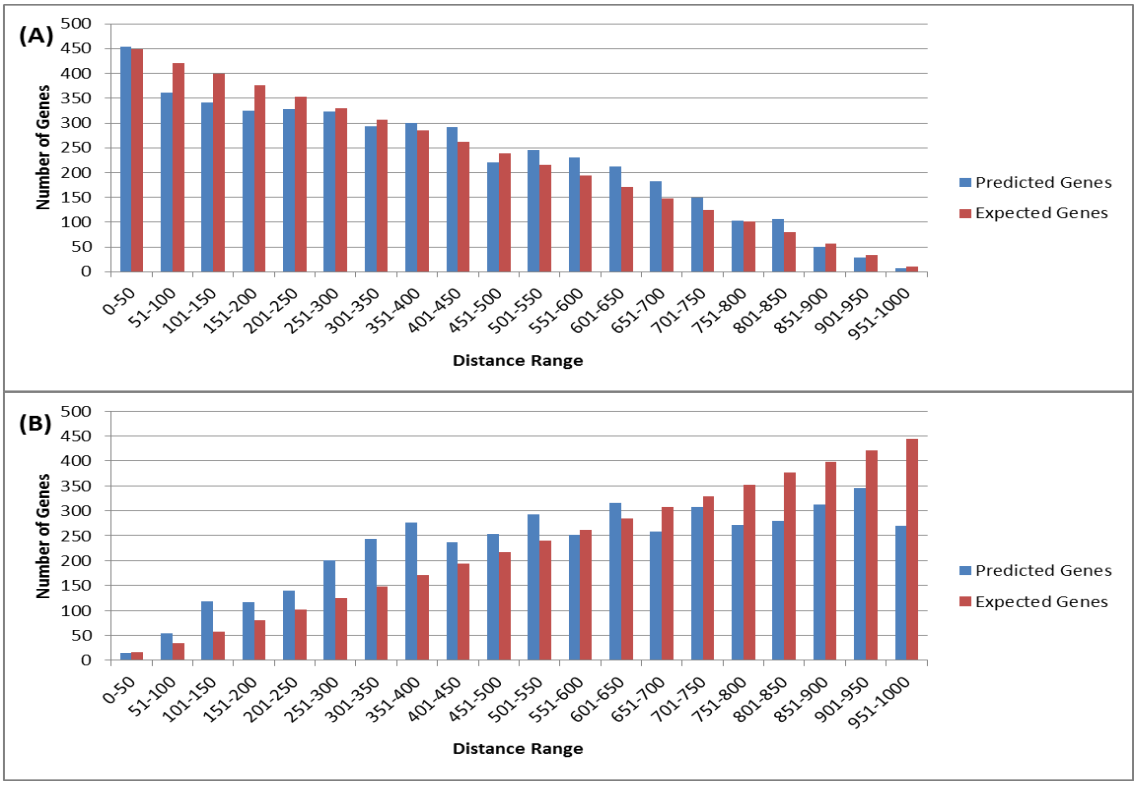


Figure 7.35: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Zn-deficiency shoots.

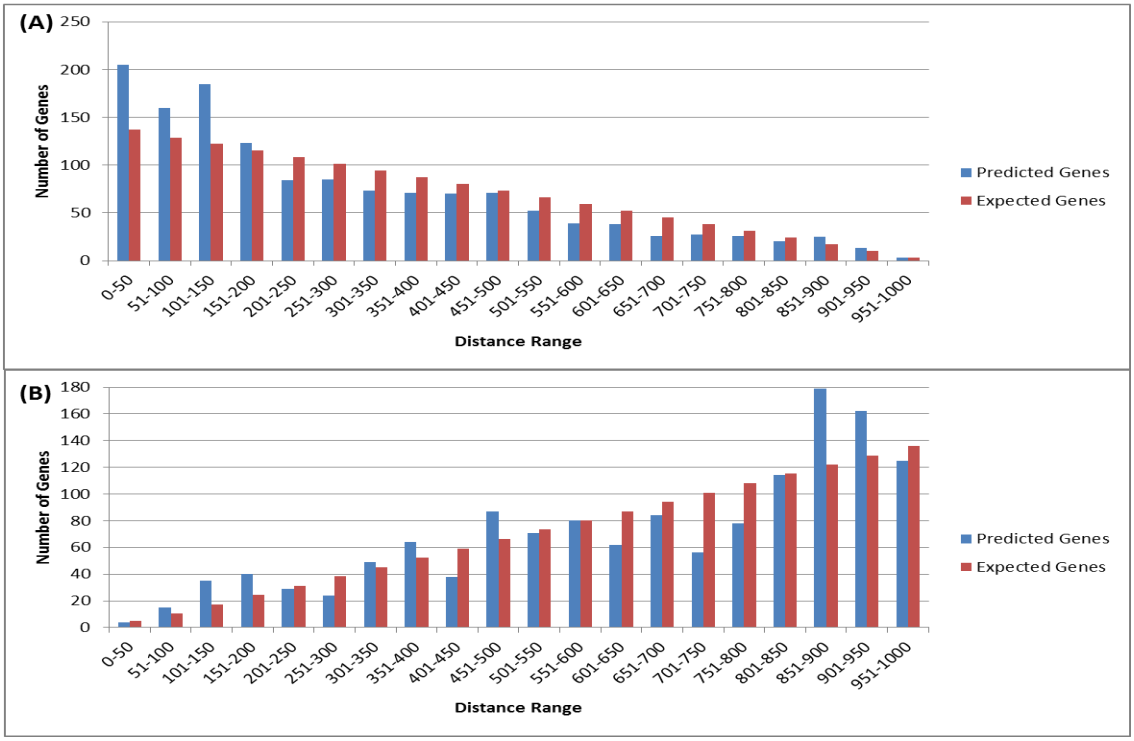


Figure 7.36: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Zn-resupplied 2h roots vs. deficient Zn shoots.

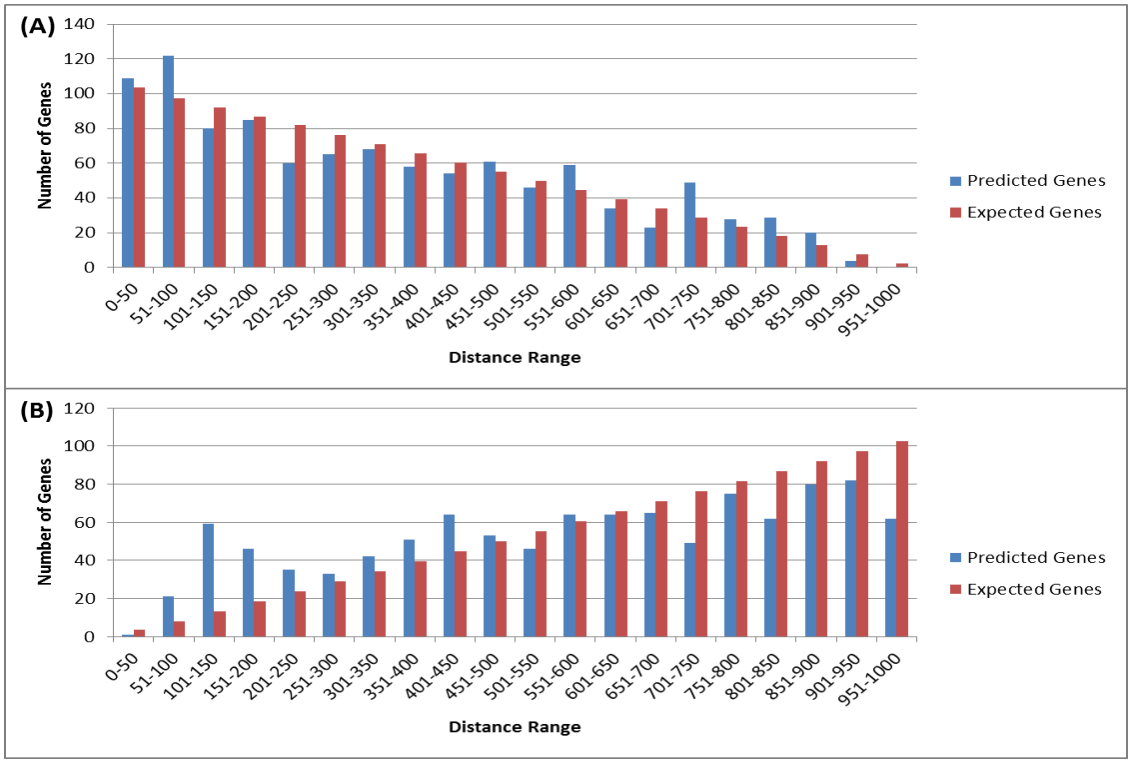


Figure 7.37: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Zn-resupplied 2h roots vs. sufficient Zn roots.

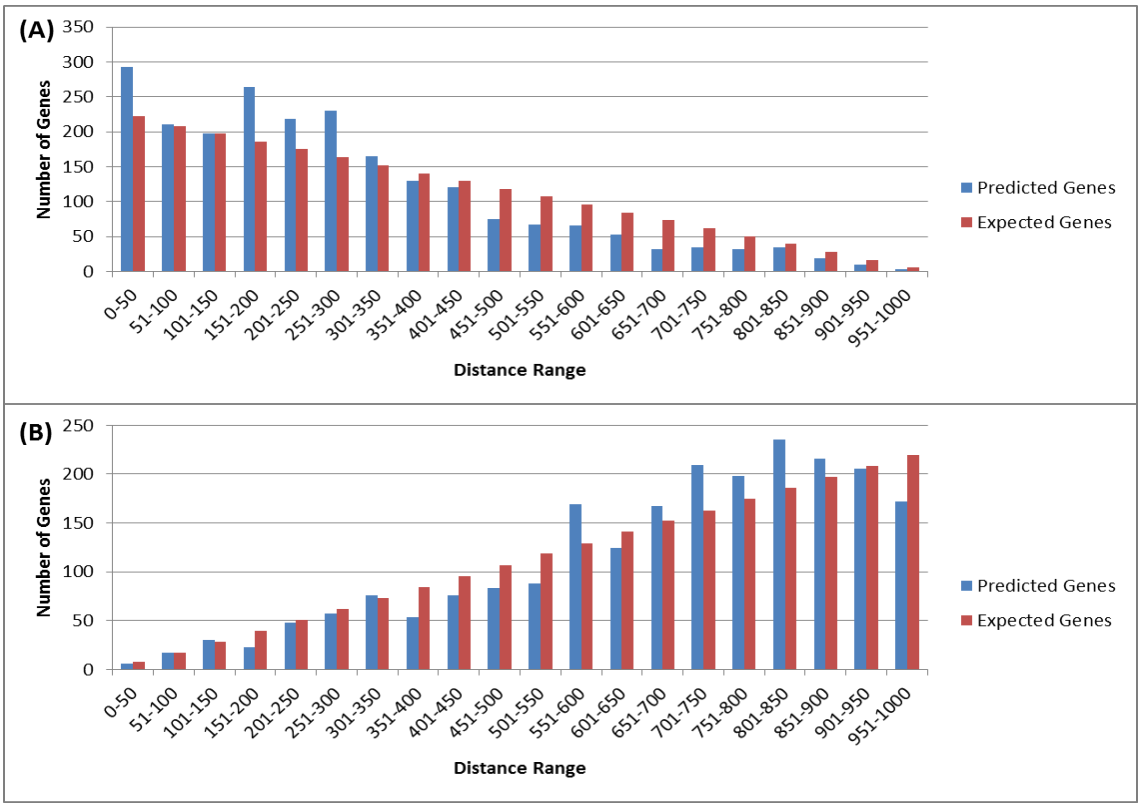


Figure 7.38: Distribution of distances to the TSS from the nearest (A) and farthest (B) motifs forming combinatorial elements putatively responsive to Zn-resupplied 8h shoots vs. sufficient Zn roots.

7.7 Ranking of pathway crosstalks

Table 7.3: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to EF-Tu 30min. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to EF-Tu 30min.

Stress	Mean	p-value
EF-Tu 60min	2.9772	2.073E-04
EF-Tu 30min	2.5147	N.A.
Salt-stressed roots 6hr	2.0457	5.654E-04
Chitooctaose	2.0314	5.396E-06
Salt-stressed roots 3hr	1.7861	2.754E-12
Harpin Z 4hr	1.7237	6.009E-14
Cold-stressed shoots 3hr	1.6862	3.856E-13
Osmotic-stressed shoots 1hr	1.5194	2.574E-21
Salt-stressed roots 24hr	1.5182	2.993E-17
Harpin Z 1hr	1.5104	3.010E-20
…	…	…

Table 7.4: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to EF-Tu 60min. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to EF-Tu 60min.

Stress	Mean	p-value
EF-Tu 60min	2.6227	N.A.
Salt-stressed roots 6hr	1.9972	2.111E-18
EF-Tu 30min	1.9482	1.176E-36
Chitooctaose	1.8003	1.832E-45
Salt-stressed roots 3hr	1.6509	2.192E-61
Harpin Z 4hr	1.6051	3.720E-73
Salt-stressed roots 24hr	1.5144	2.078E-72
Salt-stressed roots 12hr	1.4964	5.893E-83
Harpin Z 1hr	1.4810	5.365E-94
Cold-stressed shoots 3hr	1.4755	1.756E-80
…	…	…

Table 7.5: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Pb-oversupplied (50ppm) leaves. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Pb-oversupplied (50ppm) leaves.

Stress	Mean	p-value
Pb-oversupplied (50ppm) leaves	3.1079	N.A.
Pb-oversupplied (25ppm) leaves	1.7909	3.340E-61
Pb-oversupplied (50ppm) roots	1.2928	1.739E-108
Pb-oversupplied (25ppm) roots	1.2193	1.697E-141
Osmotic-stressed shoots 24hr	1.0548	5.658E-250
Osmotic-stressed shoots 12hr	1.0477	1.553E-257
Osmotic-stressed shoots 6hr	1.0342	3.568E-272
Osmotic-stressed roots 24hr	1.0266	2.225E-293
B. cinerea 18hpi	1.0243	3.384E-296
Chitin 1hr	1.0232	3.592E-171
…	…	…

Table 7.6: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Pb-oversupplied (50ppm) roots. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Pb-oversupplied (50ppm) roots.

Stress	Mean	p-value
Pb-oversupplied (25ppm) roots	2.9857	5.000E-01
Pb-oversupplied (50ppm) roots	2.7202	4.688E-03
Pb-oversupplied (50ppm) leaves	1.1346	1.889E-173
Pb-oversupplied (25ppm) leaves	1.1225	4.169E-224
Chitin 6hr	1.0630	8.093E-231
Chitin 1hr	1.0519	1.695E-223
Salt-stressed roots 6hr	1.0383	N.a.N.
Osmotic-stressed roots 12hr	1.0353	N.a.N.
Osmotic-stressed shoots 12hr	1.0350	N.a.N.
Osmotic-stressed roots 24hr	1.0346	N.a.N.
…	…	…

Table 7.7: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Pb-oversupplied (50ppm) roots. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Pb-oversupplied (50ppm) roots.

Stress	Mean	p-value
Pb-oversupplied (50ppm) roots	3.3162	N.A.
Pb-oversupplied (25ppm) roots	1.8353	4.214E-111
Salt-stressed roots 6hr	1.1094	N.a.N.
Salt-stressed roots 12hr	1.0807	N.a.N.
Salt-stressed roots 24hr	1.0797	N.a.N.
Osmotic-stressed shoots 24hr	1.0692	N.a.N.
Osmotic-stressed roots 6hr	1.0644	N.a.N.
Osmotic-stressed shoots 12hr	1.0630	N.a.N.
Pb-oversupplied (50ppm) leaves	1.0625	N.a.N.
B. cinerea 48hpi	1.0600	N.a.N.
…	…	…

Table 7.8: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-deficient shoots. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-deficient shoots.

Stress	Mean	p-value
Zn-deficient shoots	2.4408	5.000E-01
Zn-resupplied shoots 8hr vs. suffficient Zn	1.9648	7.367E-04
Zn-deficient roots	1.6451	4.037E-08
Zn-resupplied roots 2hr vs. suffficient Zn	1.3599	2.244E-16
P. syringae pv. phaseolicola 24hpi	1.2557	3.422E-19
P. syringae pv. tomato avrRpm1 24hpi	1.2339	5.970E-19
P. syringae pv. tomato 24hpi	1.2250	3.161E-19
Harpin Z 4hr	1.2207	6.847E-14
P. infestans 6hpi	1.2138	7.158E-10
Pb-oversupplied (50ppm) roots	1.2049	1.095E-14
…	…	…

Table 7.9: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-oversupplied roots 2hr. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-oversupplied roots 2hr.

Stress	**Mean**	**p-value**
Zn-oversupplied roots 2hr	2.3137	N.A.
Zn-oversupplied roots 8hr	1.8633	5.893E-02
E. orontii 5dpi	1.5402	1.593E-03
Salt-stressed shoots 1hr	1.5249	2.068E-04
P. syringae pv. maculicola avrRpt2-16hpi	1.4854	9.210E-03
ABA 24hr (30µM)	1.4689	2.472E-06
Flg22 (P. syringae) 1hr	1.4470	9.126E-05
Methyl-jasmonate 1hr	1.4248	1.269E-06
P. infestans 24hpi	1.3956	3.765E-06
Harpin Z 4hr	1.3705	1.267E-04
…	…	…

Table 7.10: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-oversupplied roots 8hr. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-oversupplied roots 8hr.

Stress	**Mean**	**p-value**
Zn-oversupplied roots 8hr	2.7211	N.A.
Zn-oversupplied roots 2hr	2.3070	1.069E-01
P. syringae pv. maculicola avrRpt2-16hpi	2.0795	1.259E-01
E. orontii 5dpi	1.8781	2.829E-02
P. syringae pv. maculicola 16hpi	1.7749	2.406E-02
Salt-stressed shoots 1hr	1.6923	1.092E-02
Harpin Z 4hr	1.6858	4.920E-03
Drought-stressed roots 6hr	1.6584	1.605E-04
Lipopolysaccharide 1hr	1.6516	3.378E-03
P. infestans 24hpi	1.6382	4.383E-04
…	…	…

Table 7.11: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-oversupplied shoots 8hr. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-oversupplied shoots 8hr.

Stress	Mean	p-value
Zn-oversupplied shoots 8hr	2.2809	N.A.
Zn-resupplied shoots 8hr vs. suffficient Zn	1.2151	4.103E-19
P. syringae pv. maculicola 16hpi	1.2001	4.268E-14
E. orontii 5dpi	1.1751	9.916E-20
Cold-stressed shoots 0.5hr	1.1637	9.153E-24
Chitin 24hr	1.1494	2.170E-16
P. syringae pv. maculicola avrRpt2-16hpi	1.1442	9.193E-15
Zn-resupplied shoots 8hr vs. deficient Zn	1.1381	2.253E-22
P. infestans 6hpi	1.1260	5.652E-20
P. syringae pv. maculicola avrRpt2-24hpi	1.1185	1.135E-18
…	…	…

Table 7.12: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-resupplied roots 2hr vs. deficient Zn. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-resupplied roots 2hr vs. deficient Zn.

Stress	Mean	p-value
Zn-resupplied roots 2hr vs. deficient Zn	2.3915	5.000E-01
Pb-oversupplied (25ppm) roots	2.1873	3.211E-01
Zn-resupplied roots 2hr vs. sufficient Zn	1.9728	3.238E-03
Salt-stressed roots 6hr	1.3990	2.281E-10
P. syringae pv. tomato hrcC- 2hpi	1.3832	8.454E-11
Salt-stressed roots 24hr	1.3549	6.292E-16
P. syringae pv. tomato avrRpm1 2hpi	1.3465	1.727E-13
P. syringae pv. tomato 24hpi	1.3325	1.286E-07
Chitin 3hr	1.3217	4.941E-09
P. syringae pv. tomato avrRpm1 6hpi	1.3155	4.182E-15
…	…	…

Table 7.13: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-resupplied roots 2hr vs. sufficient Zn. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-resupplied roots 2hr vs. sufficient Zn.

Stress	Mean	p-value
Zn-resupplied roots 2hr vs. sufficient Zn	2.2868	5.000E-01
Zn-resupplied roots 2hr vs. deficient Zn	1.8674	2.323E-03
Pb-oversupplied (25ppm) roots	1.8442	5.625E-02
Salt-stressed roots 24hr	1.3489	4.235E-20
Salt-stressed roots 6hr	1.3250	7.835E-14
P. syringae pv. tomato avrRpm1 24hpi	1.3237	4.179E-16
Osmotic-stressed shoots 24hr	1.3157	9.643E-12
P. syringae pv. tomato 24hpi	1.3149	1.798E-17
Osmotic-stressed shoots 12hr	1.3147	4.931E-15
P. syringae pv. tomato avrRpm1 6hpi	1.3117	3.636E-21
…	…	…

Table 7.14: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-resupplied shoots 8hr vs. deficient Zn. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-resupplied shoots 8hr vs. deficient Zn.

Stress	Mean	p-value
Zn-resupplied shoots 8hr vs. deficient Zn	2.8159	5.000E-01
P. syringae pv. maculicola 48hpi	1.7741	8.375E-02
Osmotic-stressed roots 0.5hr	1.6111	6.655E-04
Osmotic-stressed shoots 0.5hr	1.5302	1.119E-03
Pb-oversupplied (25ppm) roots	1.4962	1.502E-02
Cold-stressed roots 24hr	1.4608	4.474E-04
Zn-resupplied shoots 8hr vs. suffficient Zn	1.4272	2.033E-04
E. orontii 5dpi	1.3738	3.335E-04
P. syringae pv. maculicola avrRpt2-24hpi	1.3714	1.902E-03
Osmotic-stressed roots 24hr	1.3530	3.121E-04
…	…	…

Table 7.15: Top 10 stresses showing high overall mean values in crosstalk analysis for predicted CREs putatively responsive to Zn-resupplied shoots 8hr vs. sufficient Zn. Overall mean expression values are displayed on the second column. A p-value >0.05 serves to identify the most similar stresses to Zn-resupplied shoots 8hr vs. sufficient Zn.

Stress	Mean	p-value
Zn-resupplied shoots 8hr vs. sufficient Zn	2.1988	5.000E-01
Zn-deficient shoots	1.9978	1.604E-02
P. syringae pv. phaseolicola 24hpi	1.2788	3.622E-15
Zn-resupplied roots 2hr vs. suffficient Zn	1.2780	3.673E-14
Pb-oversupplied (25ppm) roots	1.2712	1.639E-09
B. cinerea 18hpi	1.2474	2.561E-16
Zn-deficient roots	1.2381	2.963E-15
P. syringae pv. tomato hrcC- 24hpi	1.2313	9.033E-20
P. infestans 6hpi	1.2310	8.353E-17
P. infestans 24hpi	1.2262	3.524E-23
…	…	…

7.8 Venn diagrams with abiotic CREs

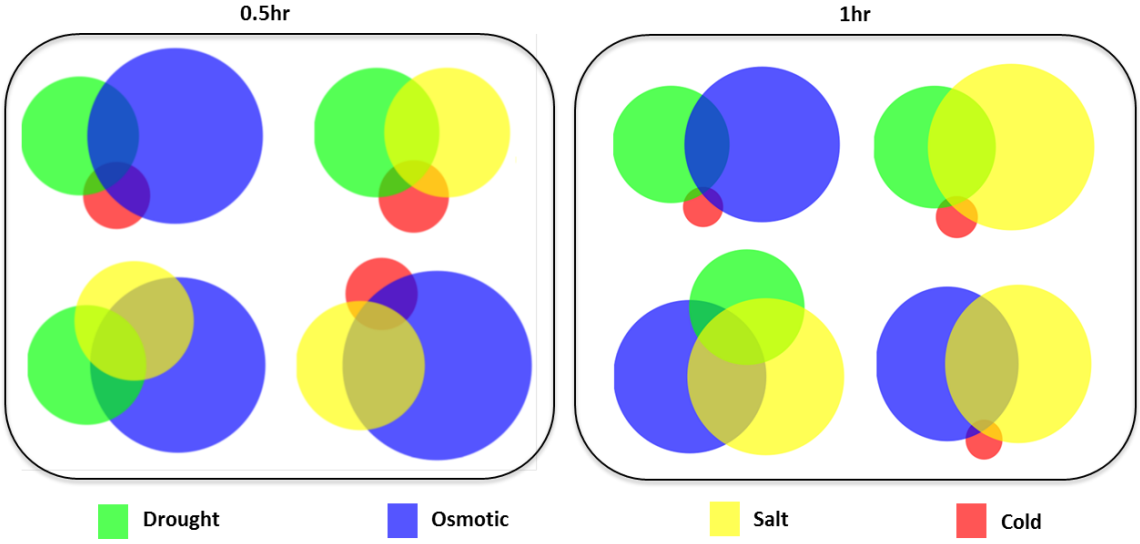


Figure 7.39: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 0.5hr and 1hr Roots putative responsive sets. Each circle represents a CRE set.

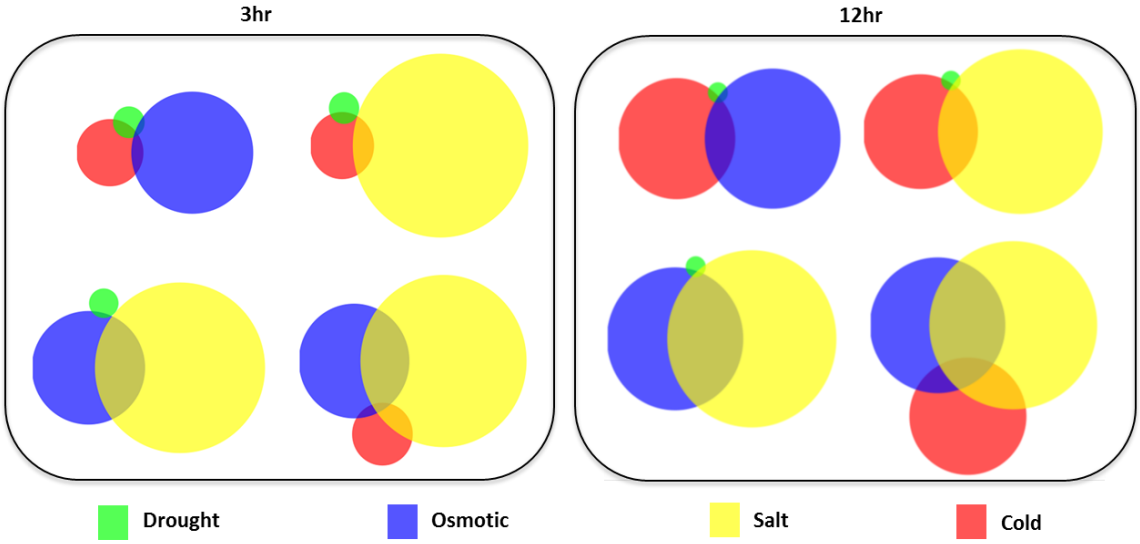


Figure 7.40: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 3hr and 12hr Roots putative responsive sets. Each circle represents a CRE set.

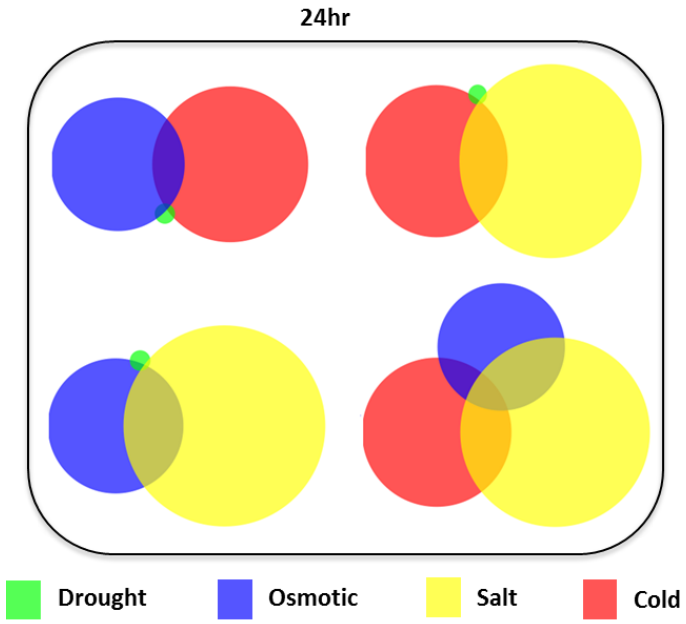


Figure 7.41: Area proportional Venn diagram showing a sequence comparison of Cold, Drought, Osmotic and Salt-stress 24hr Roots putative responsive sets. Each circle represents a CRE set.

Acknowledgements

First and foremost, I want to thank Professor Reinhard Hehl for giving me the opportunity to finish my PhD at his institute. Not only did I learn a lot with him, but his constructive criticism and advice also helped me to improve myself and to finish this work.

I also want to express my gratitude to Professor Dieter Jahn for accepting to be a co-referent of this work.

Another very important person, who supervised all of this work, was Dr. Lorenz Bülow. All the discussions we had and the ideas we developed were the basis for my work. Thank you for all the time you invested in helping me and correcting this thesis.

I also want to thank everyone at the Braunschweig Institute of Genetics. They made every day at work not only productive but also fun. I want to thank Yuri Brill for his technical support and the kind work atmosphere. I am especially grateful to Fabian. Thank you for your support at work and our friendship.

My personal gratitude goes to my family. My parents Mariano and Esperanza have supported me through all these years that I have lived abroad and were my inspiration to finish the PhD. To my brother Juan Carlos and my sisters Angelica and Carolina, thank you for always being there for me, regardless of the distance. And finally, my gratitude goes to my girlfriend Maike. She gave me all the emotional support I needed to accomplish this work.

To all of you, Dankeschön, Thank you and Gracias!